소지역통계 추정법
2차 연구결과 자료

소 지 역 통 계 추 정 법 (Ⅱ)

- 광주 · 충북의 시 · 군 · 구별 추정량분석 중심으로 -

2001. 12

통 계 기 획 국
조 사 관 리 과

시군구 실업통계 작성의 실증조사 연구

(광주광역시와 충청북도의 표본조사구를 추가한

경제활동인구특별조사의 소지역 추정법 사례연구)

2001. 12

공 군 사 관 학 교

전 산 통 계 학 과

이 계 오 교수

< 목 차>

제 1 장 개 요

현재의 경제활동인구조사는 시·도 단위까지 매월 고용동향 등을 포함한 노동력에 관한 사항의 통계를 작성하여 발표할 목적으로 1997년 말에 표본을 개편하여 실행되고 있다. 그러나 1995년부터 지방자치제도의 시행으로 시군구 단위까지 지방자치 행정이 운영되고 있으므로 각 시군구의 지방 자치단체에서는 해당 지방의 특성과 상황 여건에 적합한 지역 발전계획 수립과 효과적인 지방행정 업무를 실천하기 위해서 시군구 단위까지의 지역 통계의 생산을 요구하고 있다. 그러나 아직까지 통계생산의 제도적인 측면에서와 통계 생산기법의 이론적인 측면에서 이에 대한 해결방안을 마련하지 못하고 있다. 특히 IMF 상황에서 폭발적으로 늘어가는 실업자를 구제하기 위하여 십여 조원이상의 실업 구제대책 특별 예산을 고용증대와 공공근로사업 등에 투입하였으나 예산 집행의 효율성 측면에서 많은 문제점들이 제기되고 있다.

시·도별 예산 배정은 경제활동인구조사를 통해서 생산되는 실업자 수와 경제활동인구의 통계 수치를 기준으로 객관적이고 공평하게 배분할 수 있음에도 불구하고 각 시도에서 예하 시군구의 지방자치단체에 대한 예산 배분은 주먹구구식으로 할 수밖에 없는 실정이다. 1998년 말 정기 국회에서 실업자 통계에 대한 신뢰성에 의혹이 제기되기도 하고 모 일간지에서는 '99년 2월초에 통계 제도의 현황과 문제점을 지적하는 특집을 발간하여 신뢰성과 시의성을 갖춘 정부통계를 생산할 수 있는 통계제도와 새로운 통계기법의 개발에 대한 기반과 여건을 조성할 수 있는 계기가 되기도 하였다.

정보화 시대에서 정확하고 신속한 통계생산의 필요성이 절실해짐에 따라 실용적이고 효율적인 통계제도에 대한 연구가 추진되는등 다양한 노력들이 있었지만 시군구단위의 실업 관련 통계 생산에 대한 심층적인 연구가 미진하므로 국가적인 차원에서 이에 관한 연구가 진행되지 않는다면 몇몇 학자

들과 통계청에서 최신 정보수집과 기초적인 수준에서 이론적이고 형식적인 내용에만 관심이 집중될 뿐 실무에 활용할 수 있는 실질적인 연구는 기대하기 어려울 것이다. 행정자치부에서도 무모하게 독자적으로 전국의 시군구의 실업자 추정을 위해서 수 억원 정도의 실업 대책 예산을 낭비하면서 시도별로 표본조사를 실시했으나 엉뚱한 수치만 얻었을 뿐 신뢰성 있는 시군구의 실업통계는 얻지 못하고 아까운 국민의 세금만 낭비하였던 적이 있으며 또 일부 시도에서는 경제활동인구조사와 같은 내용의 조사를 위해서 표본 가구수를 늘려서 독자적으로 시군구단위의 실업통계의 적성을 시도했으나 신뢰할 만한 결과를 얻지못하였다.

또한 시군구의 소지역 통계 중에서 실업 관련 소지역 통계의 수요가 어느 분야보다도 많다는 사실을 조욱현과 노근호(1999)의 "소지역 통계 발전 방향"이라는 논문에서 지방자치단체의 통계실무자를 대상으로 한 설문조사를 통해서 밝히고 있다.

우리 나라와 유사한 방법으로 노동력 조사를 실행하고 있는 미국에서는 전국적으로 50,000여 표본가구를 매월 조사하여 County들을 포함하여 5,000여 종류의 소지역 실업 통계를 생산하고 있음을 감안할 때, 현재 우리 나라에서도 경제활동인구조사를 위해서 30,000여 표본가구를 매월 방문조사하고 있으므로 230여 시군구 단위에 대한 소지역 실업통계 생산이 가능할 것이나, 표본 조사된 정보만을 이용해서는 신뢰성 있는 시군구의 실업통계 생산은 힘들 것이다. 대부분의 통계 선진국들이 적용하고 있는 간접 추계방법인 소지역 추정법을 활용한다면 현재 규모의 조사원과 예산 하에서도 정책 입안 자료로 활용할 수 있을 정도의 신뢰성을 갖춘 시군구의 실업통계의 생산은 가능할것으로 생각되어 소지역 추정법을 이용하여 시군구의 실업통계 작성 기법 개발에 대해서 살펴보겠다.

먼저 간접추계법인 소지역 추정법의 개념과 기법들 중에서 실업통계관련

내용을 요약 설명하고, 우리나라의 시군구 실업통계 생산에 적용할 수 있는 모형들을 제안하겠다. 제안한 소지역 추정법의 실제 적용 가능성과 타당성을 직접현장조사를 통해서 검토 분석하기 위해서 광주광역시와 충청북도를 대상으로 실증조사를 실시하는 과정을 설명하고, 다음으로는 조사된 자료를 분석하여 시군구별 실업통계를 생산한 다음 이들을 직접추정법과 비교하고 문제점과 발전방안을 언급하여, 앞으로 2000년 인구주택총조사 자료를 근거로 2002년에 경제활동인구조사를 위한 표본개편시 반영할 사항들을 언급할 것이다. 마지막으로 결언에서 정책적인 제안과 발전방향에 대해서 요약하였으며, 시군구 실업통계작성 프로그램을 C언어로 작성하여 부록에 첨부하여 앞으로 실용적이고 심층적인 연구진행에 참고가 되도록 하였다.

제 2 장 소지역 추정법

미국이나 캐나다 등의 통계 선진국에서는 센서스의 중간 년도에 해당되는 해의 주(state) 또는 county의 인구 추정과 노동력 통계를 생산하기 위해서 전국적인 대규모의 표본조사를 실시하고 있으나 표본을 설계할 당시에는 대지역 단위의 통계를 생산할 목적이었으나 지방자치정부의 요청이나 중앙정부의 예산 배정을 위해서 좁은 지역단위의 통계를 생산할 필요가 있는 경우에는 직접 조사된 자료만을 이용한다면 신뢰성 있는 통계 생산이 불가능하므로 행정보고 자료 또는 센서스 자료 구조의 특성을 이용하여 신뢰성을 높일 수 있는 간접추계법인 소지역 추정법을 활용하고 있다.

직접 조사된 자료와 모집단의 특성과 구조 분석을 통해서 계산한 예측값을 혼합하여 표본수가 적은 점을 보완함으로써 신뢰성을 높일 수 있는 추정법을 소지역 추정법(small area estimation)이라 한다. 경제활동 인구조사에서 시도단위의 실업통계를 생산하기 위해서 시도단위까지는 충분히 많은 수의 표본조사구을 배정하였다. 그러나 시군구의 실업자를 추정하고자 한다면 각 시군구에 대해서 불균형적으로 표본이 배정될 수 있어서 어떤 시군구에는 극히 적은 수의 표본이 조사되고 또다른 시군구에는 많은 수의 표본이 조사되어 모든 시군구에서 안정되고 신뢰성 있는 추정값을 구할 수 없기 때문에 이를 해결하기 위해서는 소지역 추정법이 적용되어야 하고, 또한 시도단위에서도 성별-교육정도별로 실업자의 특성을 분석하고자 할 경우에도 각 셀(Cell)의 실업자 추정은 조사된 자료만을 이용한 직접 추정법 보다는 소지역 추정법의 적용이 타당할 것이다. 통계 분석 단위에서 인구 사회학적 특성과 성별에 의해서 세분화가 이루어 질 경우에는 통계 분석단위 내에서 세분화에 대한 구조 분석을 통해서 예측한 추정값과 조사된 자료를 함께 이용한 소지역 추정법이 적용될 수 있다.

우리 나라에서도 노동력 통계를 작성하기 위해서 5년 주기의 인구 주택 총조사에서 전체조사구의 10%를 표본조사하여 이 자료를 활용하여 설계한 경제활동인구조사의 표본틀을 현행대로 유지하면서 모집단의 구조분석과 노동부의 중앙고용정보원의 구직등록 데이터베이스의 자료와 행정자치부의 주민등록인구통계를 이용하는 소지역 추정법을 시군구의 실업통계 생산에 적용할 수 있는 추정법에 대한 연구를 살펴보자. 표본조사구가 충분하게 배분된 시군에서는 현재의 직접 추정법을 적용할 수있는지를 분석하기위해서 직접추정법을 먼저 설명하고 간접 추정법과 모형기반 추정법을 가능한 자세하게 설명하겠다.

2.1 직접 추정법(Direct Estimation)

통계청에서는 매달 약 30,000표본 가구에 대해 경제활동인구조사를 실시하고 있다. 국가의 고용현황 및 경제 정책 등은 이러한 월별 조사 결과를 토대로 작성된다. 통계청의 경제활동인구조사에 대한 관심영역은 7개 광역시와 9개 도 단위를 포함한 16개의 대영역이다.

이러한 지역들에 대한 실업자 총계를 추정하기 위한 직접추정량은 다음과 같은 총계 추정공식이 이용된다.

$$\hat{Y}_{i\cdot} = \sum_{s=1}^{2} {}_s\hat{Y}_{i\cdot} \quad , \; i=1,2,\cdots,I \; ; s=1,2 \; ; h=1,2,\cdots,n_i$$

$$= \sum_{s=1}^{2}\sum_{h=1}^{n_i} {}_s\hat{Y}_{ih}$$

$$= \sum_{s=1}^{2}\sum_{h=1}^{n_i} {}_sM_i \; {}_sY_{ih} \qquad (2.1)$$

여기에서 s는 성별(남-여)을 나타내는 첨자, n_i는 경제활동인구조사에서 i번째 지역의 표본조사구 수, ${}_sY_{ih}$는 각 성별에 대해서 i번째 지역의 h번째 표본 조사구에서 조사한 실업자 수를 나타낸다. 승수 ${}_sM_i = {}_s\hat{X}_{i\cdot} / {}_sX_{i\cdot}$ 은 $\hat{Y}_{i\cdot}$ 이 불편추정량이 되도록 산정한다. 여기에서 ${}_s\hat{X}_{i\cdot}$ 은 i번째 지역에 대한 15세 이상의 상주 추계인구를 나타내며, ${}_sX_{i\cdot}$ 는 경제활동인구조사에서

조사된 15세 이상의 조사인구를 나타낸다.

직접추정량 $\hat{Y}_{i\cdot}$의 분산은 다음 (2.2)식과 같이 주어진다.

$$Var(\hat{Y}_{i\cdot}) = \sum_{s=1}^{2} Var({}_s\hat{Y}_{i\cdot}) + 2\,Cov({}_1\hat{Y}_{i\cdot}, {}_2\hat{Y}_{i\cdot}), \quad i=1,2,\cdots,I$$

$$= \sum_{s=1}^{2} {}_sM_i^2\, Var(\sum_{h=1}^{n_i} {}_sY_{ih}) + 2\,{}_1M_i\,{}_2M_i\,Cov(\sum_{h=1}^{n_i} {}_1Y_{ih}, \sum_{h=1}^{n_i} {}_2Y_{ih}). \quad (2.2)$$

$\hat{Y}_{i\cdot}$의 분산에 대한 추정값은 다음 (2.3)식의 추정공식을 이용하여 계산된다.

$$\hat{Var}(\hat{Y}_{i\cdot}) = \sum_{s=1}^{2} {}_sM_i^2(\zeta_i \sum_{h=1}^{n_i} {}_sU_{ih}^2) + 2\,{}_1M_i\,{}_2M_i(\zeta_i \sum_{h=1}^{n_i} {}_1U_{ih}\,{}_2U_{ih}), \quad (2.3)$$

여기에서 ${}_sU_{ih} = d\,{}_sY_{ih} - {}_s\rho_i \cdot d\,{}_sX_{ih}$, $d\,{}_sY_{ih} = {}_sY_{ih} - {}_sY_{i,h+1}$,

$d\,{}_sX_{ih} = {}_sX_{ih} - {}_sX_{i,h+1}$, ${}_s\rho_i = {}_sY_{i\cdot} / {}_sX_{i\cdot}$,

$\zeta_i = [1-n_i/(10N_i)]\,n_i/[2(n_i-1)]$ 이고, N_i는 소지역 i에 대한 모집단의 조사구 수를 나타낸다.

경제활동인구조사에서 대영역에 포함된 소지역들은 표본설계를 계획할 당시에 반영된 관심영역이 아니다. 따라서 대영역 표본설계에 기반을 둔 표본 조사로부터 소지역들에 대한 관심 통계량들을 추정한다면 소지역에 할당된 표본 조사구 수가 충분하지 않기 때문에 신뢰할 만한 결과를 얻을 수 없다. 이러한 관점에서 볼 때 경제활동인구조사에서 소지역에 대한 직접추정값은 목표 정도를 만족할 수 없는 경우가 많기 때문에 이를 보완하는 방법으로서 행정업무보고자료 또는 통계모형의 가정을 설정하고 예측값을 계산하여 이용하는 간접추정법이 이용될 수 있으므로 이에 대한 일반적인 사항에서부터 알아보자.

2.2. 간접 추정법(Indirect Estimation)

2.2.1 인구통계적 방법(Demographic Method)

미국의 경우에서와 같이 10년 주기로 센서스를 할 경우, 지방 도시나 카운티(county)의 중간 해당 년도의 인구를 추정하기 위해서 사용하는 추정법으로 센서스 자료와 인구수에 관련된 징후 변수(출생자수, 사망자수, 주택 수, 등록한 학생 수 등)의 변동을 분석하여 얻은 예측값을 결합하는 추정법을 인구 통계적 방법이라 말한다.

(1) 생존률법(Vital Rates Method: VR Method)

VR법은 출생과 사망에 관련된 자료를 이용하여 인구의 변동률보다는 징후 변수의 영향만을 분석하여 활용한다. 가장 최근에 센서스를 실시한 해를 기준 년도로 하고, 기준해로부터 t년 후에 소지역의 인구수를 추계하고자 한다. 여기에서 전제 조건은 추계 대상인 소지역을 포함하는 대지역의 특성과 소지역의 특성이 동일하다는 것이며, 전제 조건에서 많이 벗어나는 경우에는 추정량의 편향이 커져서 신뢰도가 낮아진다.

t년 후의 소지역의 출생률과 사망률을 γ_{bt}와 γ_{dt}로 표현하고 대지역의 출생률과 사망률을 R_{bt}와 R_{dt}라 나타내면 다음과 같은 관계가 주어진다.

$$\gamma_{bt} = \gamma_{bo}\left(\frac{R_{bt}}{R_{bo}}\right), \qquad \gamma_{dt} = \gamma_{do}\left(\frac{R_{dt}}{R_{do}}\right) \tag{2.4}$$

여기에서 γ_{b0}와 γ_{d0}는 기준 해의 소지역의 출생률과 사망률이고 R_{b0}와 R_{d0}는 기준 해의 대지역의 출생률과 사망률을 의미한다.

센서스를 실시한 기준해로부터 t년 후의 인구수는 다음 식에 의해서 추계할 수 있다.

$$p_t = \frac{1}{2}\left(\frac{b_t}{\gamma_{bt}} + \frac{d_t}{\gamma_{dt}}\right) \tag{2.5}$$

단, b_t와 d_t는 소지역의 t년 후의 출생자수와 사망자수를 뜻한다.

(2) 성분법(Components Method)

성분법은 출생과 사망 인구수 및 유입, 유출 인구에 관한 자료를 이용하여 소지역의 인구 수를 추정하기 위해 고안된 방법이다.

센서스를 실시한 기준해로부터 t 년 동안의 출생 인구, 사망인구 및 총 이주인구를 각각 $b_{0,t}$, $d_{0,t}$, $m_{0,t}$ 로 나타냈을 때 t년 후의 인구수는 다음식에 의해 추정한다.

$$p_t = p_0 + b_{0,t} - d_{0,t} + m_{0,t} \qquad (2.6)$$

여기에서 $m_{0,t} = i_{0,t} - e_{0,t} + n_{0,t}$ 로 계산하며, $i_{0,t}$ 는 유입인구, $e_{0,t}$ 는 유출인구, $n_{0,t}$ 는 주(state)간의 총 이주인구를 나타내며 행정보고자료에 의해 주어진다.

(3) 회귀 징후법(Regression Symptomatic Procedures)

회귀 징후법은 다중선형회귀모형을 이용하여 소지역의 인구를 추정하는 방법으로써 징후변수들을 독립변수로 선택하여 소지역 추정에 이용한다. 비상관계수(Ratio Correlation), 차분 상관계수(Difference Correlation), 표본 회귀법(Sample Regression Method) 등은 이러한 회귀징후법의 일종이다. 여기에서는 다른 두 방법보다는 비교적 자주 사용되고 있는 표본 회귀법을 설명하기로 한다.

먼저 종속변수와 독립변수를 다음과 같이 정의하자.

$Y_i = (p_{it}/P_t)/(p_{i0}/P_0) =$ i 소지역의 인구비 변화량,

$x_{ij} = (s_{ijt}/S_{jt})/(s_{ij0}/S_{j0}) =$ i 소지역에 대한 j 번째 징후변수 s_j 의 변화량,

여기에서 P_t , P_0 , S_{jt} , S_{j0}는 소지역 i 를 포함하는 대지역에서의 값들이고, x_{ij} 는 행정자료로부터 얻는다($j = 1, 2, \cdots, p$).

표본 회귀법은 종속변수 Y_i 가 징후변수 $x_{i1}, x_{i2}, \cdots, x_{ip}$ 의 일차결합으로 표현될 수 있다는 것을 가정하며, 이때 Y_i 의 값은 조사된 직접추정값 $\hat{Y}_i$ 을 이용하여 m개의 소지역 중 k개의 소지역에 대하여 선형회귀식을 적합시켜 회귀계수들을 추정한 후, Y_i 의 추정값으로 다음의 표본 회귀추정량을 이용한다.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} \quad , \quad i = 1, 2, \cdots, m \qquad (2.7)$$

i 소지역에 대한 인구수는 (2.7)식의 표본 회귀추정량을 이용하여 다음식으로 추정한다.

$$\hat{p}_{it} = \hat{Y}_i (p_{i0}/P_0)\ \hat{P}_t \quad , \quad i = 1, 2, \cdots, m \qquad (2.8)$$

표본 회귀추정량은 표본으로부터 직접 추정된 값이 아니라 다중선형회귀를 거쳐 얻어진 보정된 추정량이며, 표본 회귀법은 이를 이용하여 소지역의 인구를 추정하는 방법이다. 그러나 이러한 방법은 추후 논의될 모형에 근거한 소지역 추정보다는 효율성이 상당히 떨어지는 것으로 밝혀지고 있다.

2.2.2 합성 추정법(Synthetic Estimation)

추정하고자 하는 소지역과 특성이 유사한 소지역들의 정보를 이용하여 추정값의 정도를 높이고자하는 추정방식을 합성 추정법이라하며, 주변이나 유사지역의 정보를 이용하므로 "Borrow Strength"라고 말하기도 한다. 표본조사의 설계 시에는 대영역에 대해서만 직접 추정값을 구하고자 하였으나 대영역을 분할한 소지역의 추정값이 필요한 때에는 대영역과 소지역의 구조적 특성이 같다는 조건하에서 소지역의 연구변수에 대한 추정값을 구할 수 있는데, 이때 대영역의 분할은 지리적인 분할보다는 연령대별 또는 교육정도별과 같은 특성에 따른 분할을 말한다. 우리나라의 경활조사의 예로 살펴보면 조사단위는 동부와 읍면부로 나누어져 광역단체별로 층화추출되므로 시

지역과 읍면 지역으로 그룹을 나눈다면 각 그룹내에서는 연령대별 구조나 교육정도별 구조의 특성이 거의 유사한 것으로 가정할 수 있고, 이러한 경우 시군구 실업자 추정에 합성추정법의 이용도 가능할 것으로 생각된다.

대영역을 I개 소지역으로 분할하며 또한 대영역을 특성 기준에 따라 J개의 범주로 분류한다면 i소지역의 추정값은 다음 식으로 구할 수 있다.

$$\hat{Y}_{i\cdot} = \sum_{j}^{J} p_{ij}\ \hat{Y}_{\cdot j} \tag{2.9}$$

단, p_{ij}는 i번째 소지역의 j범주에 대한 가중값이며 센서스나 행정자료에서 구해진다. $Y_{\cdot j}$는 대영역에서 j범주에 대한 표본에서 구한 추정값이다. 단 대영역의 표본의 수는 충분하게 많아서 신뢰성 있는 추정값을 구할 수 있다고 가정한다.

i소지역의 실업자 추정에 관한 경우를 생각해 보자.

Y_{ij} : i소지역의 j범주(연령대별 또는 교육정도별)의 실업자수,

X_{ij} : i소지역의 j범주(연령대별 또는 교육정도별)의 경제활동인구,

$\hat{Y}_{\cdot j} = \sum_{i} Y_{ij}$: j범주의 대영역에 대한 합계,

$Y_{i\cdot} = \sum_{j} Y_{ij}$: i소지역의 실업자 수.

$\hat{Y}_{\cdot j}$의 직접 추정값($\hat{Y}_{d\cdot j}$)은 표본조사 자료만으로 계산가능하고, X_{ij}는 센서스 또는 행정자료 등 보조변수의 정보에서 계산 가능한 것으로 가정한다면 합성추정량은 다음과 같이 나타낼 수 있다.

$$\hat{Y}_{i\cdot}^{\ s} = \sum_{j} \left(\frac{X_{ij}}{X_{\cdot j}} \right) \hat{Y}_{d\cdot j} \tag{2.10}$$

만일에 $\hat{Y}_{d\cdot j}$가 비추정량의 형식을 갖는다면, $\hat{Y}_{d\cdot j} = (\hat{Y}_{\cdot j} / \hat{X}_{\cdot j}) X_{\cdot j}$ 로 나타낼 수 있으므로 (2.10)식은 다음과 같이 표현될 수 있다.

$$\hat{Y}_{i.}^{s} = \sum_j X_{ij}\left(\frac{\hat{Y}_{.j}}{\hat{X}_{.j}}\right) = \sum_j \left(\frac{X_{ij}}{\hat{X}_{.j}}\right)\hat{Y}_{.j} \qquad (2.11)$$

여기에서 $\hat{Y}_{i.}^{s}$가 불편추정량이 되기 위해서는 $\frac{Y_{.j}}{X_{.j}} = \frac{Y_{ij}}{X_{ij}}$ 를 만족해야 하고, 이를 만족하지 못할 경우에는 편향추정량이 되고, 이때 $\hat{Y}_{i.}^{s}$의 편향(Bias)의 크기는 $B(\hat{Y}_{i.}^{s}) = E(\hat{Y}_{i.}^{s} - Y_{i.})$이다. 즉, 편향의 크기는 아래 식으로 표현될 수 있다.

$$B(\hat{Y}_{i.}^{s}) = \sum_j X_{ij}\left(\frac{Y_{.j}}{X_{.j}} - \frac{Y_{ij}}{X_{ij}}\right).$$

$\hat{Y}_{i.}^{s}$는 편향 추정량이므로 $\hat{Y}_{i.}^{s}$의 변동의 크기는 평균제곱오차로 표현되어야 하지만 이의 추정량에 대한 확정적인 형식의 표현이 없으므로, 근사적 불편추정량이 아래와 같이 주어질 수 있다.

$$\widehat{MSE}(\hat{Y}_{i.}^{s}) = (\hat{Y}_{i.}^{s} - \hat{Y}_{i.})^2 - \widehat{Var}(\hat{Y}_{i.}) \qquad (2.12)$$

2.2.3 복합 추정법(Composite Estimation)

소지역에 배정된 표본수가 적기 때문에 조사된 자료만을 이용한 직접 추정량의 불안정에서 오는 낮은 신뢰성과 합성추정량의 편향을 보완하기 위해서 직접 추정값과 합성 추정값의 가중평균을 사용할 수 있는데 이를 복합 추정량(Composite Estimator)이라 한다.

$$\hat{Y}_{i.}^{c} = w_i\,\hat{Y}_{i.} + (1-w_i)\,\hat{Y}_{i.}^{s} \qquad (2.13)$$

여기에서 $\hat{Y}_{i.}$는 조사된 자료에서 직접 계산한 추정값이며, $\hat{Y}_{i.}^{s}$는 합성 추정값을 나타낸다. w_i는 가중값으로 0과 1사이의 값이다.

이제 가중값 w_i를 산정하는 방법을 알아보자. 먼저 평균제곱오차 $MSE(\hat{Y}_{i.}^{c})$를 최소화하는 w_i는 아래와 같다.

$$w_{i(opt)} = \frac{MSE(\ \hat{Y}_{i.}^{s})}{MSE(\ \hat{Y}_{i.}^{s}) + V(\ \hat{Y}_{i.})}$$

최적 가중값 $w_{i(opt)}$의 추정값은 다음 식으로 계산된다.

$$\hat{w}_{(opt)} = \frac{mse(\ \hat{Y}_{i.}^{s})}{(\ \hat{Y}_{i.}^{s} - \hat{Y}_{i.})^2}$$

다음에는 모든 소지역에 공통 가중값을 부여하는 방법으로써 초기 공통 가중값 w를 이용하여 $MSE(\ \hat{Y}_{i.}^{s})$들의 평균을 최소화하는 가중값은 아래와 같다.

$$\hat{w}_{(opt)} = 1 - \frac{\sum_i \hat{V}(\ \hat{Y}_{i.})}{\sum_i (\ \hat{Y}_{i.}^{s} - \hat{Y}_{i.})^2} \qquad (2.14)$$

세 번째 산정법은 각 소지역에 배정된 표본 크기에 의존하는 가중값은 다음과 같이 계산된다.

$$w_i(\delta) = \begin{cases} 1, & \hat{N}_i \geq \delta N_i \\ \dfrac{\hat{N}_i}{\delta N_i}, & (\text{그외}) \end{cases} \qquad (2.15)$$

단, N_i는 i소지역의 크기이며 $\hat{N}_i = N(n_i/n)$이다. $\hat{N}_i$는 직접추정량이며 δ는 합성추정량의 기여도를 조정하는 값이므로 주관적으로 결정한 값이다. 예를 들어 캐나다 노동력 통계조사에서는 2/3으로 한다.

어떤 추정법에 의해서 소지역의 추정값을 구하더라도 대영역을 소지역으로 분할하여 각 소지역의 추정값을 추정하므로 소지역의 추정값의 합계는 대영역의 추정값과 같아야 할 것이다. 왜냐하면 매월 정부기관에서 발표하는 광역시와 도의 실업자수와 해당 소지역의 추정값의 합계가 같도록 조정하지 않으면 서로 상이한 통계수치로 인하여 혼란을 줄 수 있기 때문에 한

가지 통계수치가 되도록 조정된 추정량을 계산해야 할 것이다. 각 소지역의 추정량을 생존률법(VR Method), 합성추정법 또는 복합 추정법 중 어느 한 방법으로 계산한 것으로 간주할 때 조정된 소지역 추정량은 다음과 같다.

$$\hat{Y}_{i.}^{A} = \left(\frac{\hat{Y}_{i.}^{*}}{\sum_{i} \hat{Y}_{i.}^{*}}\right) \hat{Y} \qquad (2.16)$$

단, $\hat{Y}$는 광역시·도의 직접 추정값이며, $\hat{Y}_{i.}^{*}$는 i소지역의 *추정법으로 추정한 것이다.

2.3 모형 기반 추정법(Model-Based Estimation)

2.3.1 기본적 지역 수준 모형(Basic Area-level Model)

소지역 추정시 모형에 근거한 추정방법이 많은 사람들의 관심을 끌고 있는 것은 다음과 같은 몇가지 장점에 기인한다. 먼저 모형 기반 추정법은 소지역들을 연결하고 있는 모형 구조가 소지역 간의 복잡한 오차구조를 내포하고 있기 때문에 소지역 간의 변동을 반영하여 소지역 추정의 정확도를 높일 수 있다는 점이며, 또한 표본자료로부터 모형의 유용성을 확인할 수 있으며, 연속형의 자료뿐만 아니라 범주형 자료 및 시계열 자료와 같은 다양한 경우들에 대해서도 모형화하여 추론할 수 있으며, 모형 기반 추정법으로 소지역 추정량들과 연관있는 많은 측도들이 얻어질 수 있다는 장점들을 들 수 있다.

지역 간의 공변량을 포함하고 있는 지역 수준 모형을 이용하여 경험적 최량선형불편예측(EBLUP) 추정량, 경험적 베이즈(EB) 추정량, 계층적 베이즈(HB) 추정량에 대해 설명하기로 한다. 지역 수준 모형은 기본적으로 두 가지의 성분들로 이루어진다. 즉, 소지역에 대한 직접추정량 $\hat{\theta}_i$과 소지역의 보조변수들로 표현되는 θ_i의 두 가지 성분들을 모형으로 연결하여 모형

기반 추정량을 찾아내게 된다.

지정된 함수 $g(\cdot)$에 대하여 직접 추정량 $\hat{\theta}_i = g(\hat{\overline{Y}}_i)$은 모집단의 값 $\theta_i = g(\overline{Y}_i)$와 표본오차 e_i에 의해 다음과 같이 표현될 수 있다.

$$\hat{\theta}_i = \theta_i + e_i \quad , \quad i = 1, 2, \cdots, m \tag{2.17}$$

여기에서 표본오차 e_i는 서로 독립이며, 평균이 0, 분산이 ψ_i 이라고 가정하며, 보통 ψ_i는 알려진 것으로 가정한다.

θ_i는 소지역의 정보를 나타내는 보조변수 $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \cdots, z_{ip})^T$를 이용하여 선형회귀모형을 통해 표현한다.

$$\begin{aligned} \theta_i &= z_{i1}\beta_1 + z_{i2}\beta_2 + \cdots + z_{ip}\beta_p + v_i \\ &= \boldsymbol{z}_i^T \boldsymbol{\beta} + v_i \end{aligned} \tag{2.18}$$

여기에서 모형오차 v_i는 서로 독립이며, 평균이 0, 분산 σ_v^2을 갖고, 표본오차 e_i와는 서로 독립임을 가정한다.

마지막으로 (2.17)과 (2.18)의 두 성분들을 결합하면 다음과 같은 결합모형을 얻을 수 있다.

$$\hat{\theta}_i = \boldsymbol{z}_i^T \boldsymbol{\beta} + v_i + e_i \tag{2.19}$$

위의 결합모형은 고정효과 $\boldsymbol{\beta}$와 소지역 랜덤효과 v_i를 갖는 선형혼합효과모형의 일종이며, 특히 설계 기반 확률변수(design-based random variable) e_i와 모형 기반 확률변수(model-based random variable) v_i를 동시에 포함하고 있는 모형이다. 여기에서 모수 σ_v^2은 소지역들의 동질성을 나타내는 측도이다.

2.3.2 경험적 최량선형불편예측(EBLUP) 방법

경험적 최량선형불편예측(EBLUP) 방법, 경험적 베이즈(EB) 방법 및 계

층적 베이즈(HB) 방법은 모형에 근거한 소지역 추정문제에 많이 활용되고 있는 방법이다. 특히 경험적 최량선형불편예측 방법은 선형혼합모형을 이용한 추론에 응용되어 왔고, 경험적 베이즈 방법 및 계층적 베이즈 방법은 좀 더 일반적인 모형을 이용한 소지역 추정에 활용되고 있다.

EBLUP 추정량은 랜덤오차 e_i와 v_i의 분포에 대한 가정을 필요로 하지 않으나, *MSE* 추정을 위해 정규분포를 가정하기도 한다. 또한, EBLUP 추정량과 EB 추정량은 e_i와 v_i를 정규분포로 가정했을 경우에는 동일하며, HB 추정량과는 근사적으로 같게 나타난다. 그러나 추정량들의 변동을 나타내는 측도들은 동일하지는 않다.

고정계수 l_t를 갖는 θ_i 의 선형추정량 $\sum l_t \hat{\theta}_t$ 가 모형 (2.19)에 대해서 $\sum l_t \hat{\theta}_t - \theta_i$ 의 기대값이 0 을 만족할 때, $\sum l_t \hat{\theta}_t$ 를 θ_i 의 선형불편예측(LUP) 추정량이라 한다. θ_i 의 최량선형불편예측(BLUP) 추정량은 선형불편예측(LUP) 추정량들 중 최소평균제곱오차를 갖는 추정량을 말한다.

모형 (2.19)하에서 θ_i 의 BLUP 추정량은 다음과 같이 주어진다(Prasad and Rao,1990).

$$\tilde{\theta}_i(\sigma_v^2) = \gamma_i \hat{\theta}_i + (1-\gamma_i)\ \boldsymbol{z}_i^T\ \tilde{\boldsymbol{\beta}}(\sigma_v^2) \qquad (2.20)$$

여기에서 $\gamma_i = \sigma_v^2/(\sigma_v^2 + \psi_i)$이고, $\tilde{\boldsymbol{\beta}}(\sigma_v^2)$은 가중치 $(\sigma_v^2 + \psi_i)^{-1}$을 갖는 가중최소제곱추정량으로 아래와 같이 주어진다.

$$\tilde{\boldsymbol{\beta}}(\sigma_v^2) = (\sum_i \gamma_i\ \boldsymbol{z}_i\ \boldsymbol{z}_i^T)^{-1}(\sum_i \gamma_i\ \boldsymbol{z}_i y_i) \qquad (2.21)$$

(2.20)식의 BLUP 추정량은 가중치 γ_i를 갖는 직접추정량 $\hat{\theta}_i$ 과 가중치 $1-\gamma_i$를 갖는 회귀합성추정량 $\boldsymbol{z}_i^T \tilde{\boldsymbol{\beta}}(\sigma_v^2)$의 가중결합으로 볼 수 있다. 또한, 표본분산 ψ_i가 작을때(σ_v^2이 클 경우) BLUP 추정량은 직접추정량

$\hat{\theta}_i$ 에 큰 가중치가 부여되고, 반대의 경우에는 회귀합성추정량 $\boldsymbol{z}_i^T \tilde{\boldsymbol{\beta}}(\sigma_v^2)$에 큰 가중치가 부여된다. 표본이 추출되지 않은 지역들에 대해서는 BLUP 추정량은 회귀합성추정량만으로 주어질 수 있다.

BLUP 추정량의 변동의 측도는 추정량의 $MSE = E(\hat{\theta}_i - \theta_i)^2$에 의해 주어지며 다음과 같다.

$$MSE\{\hat{\theta}_i(\sigma_v^2)\} = g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2) \qquad (2.22)$$

단, $g_{1i}(\sigma_v^2) = \gamma_i \phi_i$이고, $g_{2i}(\sigma_v^2) = \sigma_v^2(1-\gamma_i)^2 \boldsymbol{z}_i^T(\sum_i \gamma_i \boldsymbol{z}_i \boldsymbol{z}_i^T)^{-1} \boldsymbol{z}_i$로 주어진다. 식(2.20)과 (2.22)은 랜덤오차 v_i와 e_i에 대한 분포의 가정을 필요로 하지는 않는다.

주요 항 $g_{1i}(\sigma_v^2) = \gamma_i \phi_i$는 $O(1)$, $g_{2i}(\sigma_v^2)$은 $O(m^{-1})$의 형식 유계인 항이며, 이로부터 BLUP 추정량의 MSE 값은 γ_i 나 모형분산 σ_v^2이 표본분산 ϕ_i 에 비해 작을 경우에는 직접추정량의 MSE 값보다 훨씬 작아질 수 있다는 사실을 알 수 있다. 따라서 소지역 추정의 정확도는 표본분산에 비해 모형분산을 작게할 수 있는 보조변수에 크게 의존한다고 볼 수 있다.

대부분의 문제에서는 모형분산 σ_v^2은 미지인 값이므로 적절한 추정값 $\hat{\sigma}_v^2$을 산정하여 EBLUP 추정량 $\hat{\theta}_i = \hat{\theta}_i(\hat{\sigma}_v^2)$을 산출한다. 이때 소지역의 평균 $\overline{Y}_i$ 의 추정량($\hat{\overline{Y}}_i$)은 $g^{-1}(\hat{\theta}_i)$ 로, σ_v^2의 추정량은 $\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$로 주어진다. 여기에서 $\tilde{\sigma}_v^2$은 다음 식을 만족한다.

$$(m-p)\,\tilde{\sigma}_v^2 = \sum_i (\hat{\theta}_i - \boldsymbol{z}_i^T \boldsymbol{\beta}^*)^2 - \sum_i \phi_i h_{ii} \qquad (2.23)$$

(2.23)식에서 $h_{ii} = \boldsymbol{z}_i^T(\sum_i \boldsymbol{z}_i \boldsymbol{z}_i^T)^{-1} \boldsymbol{z}_i$ 이고, $\boldsymbol{\beta}^*$는 $\boldsymbol{\beta}$의 OLS(ordinary least squares) 추정량이다. 한편, $\tilde{\sigma}_v^2$은 아래와 같은 비선형 방정식의 반

복적인 해로써 구할 수도 있다.

$$a(\sigma_v^2) = \sum_i \{\hat{\theta}_i - z_i^T \tilde{\beta}(\sigma_v^2)\}^2 / (\sigma_v^2 + \phi_i) = m - p \qquad (2.24)$$

여기에서 $\tilde{\beta}(\sigma_v^2)$은 (2.21)식에 주어졌고, (2.24)식의 가운데 항은 가중잔차제곱합, $m-p$ 는 가중잔차제곱합과 관계 있는 자유도이다. 만약 $\hat{\sigma}_v^2 = 0$ 이면 EBLUP 추정량 $\hat{\theta}_i$ 는 회귀합성추정량 $z_i^T \hat{\beta}$ 로 축약된다. 단, $\hat{\beta} = \tilde{\beta}(\hat{\sigma}_v^2)$이며, 식(2.21)에서 σ_v^2 대신에 $\hat{\sigma}_v^2$을 대체하여 산출한다. 물론 위의 (2.23), (2.24)식으로부터 얻게되는 추정량들도 v_i와 e_i의 분포에 대한 가정을 필요로 하지는 않는다.

만약 랜덤오차 v_i와 e_i가 정규분포를 따른다고 가정한다면, $\hat{\theta}_i$는 평균이 $z_i^T \beta$ 이고 분산이 $\sigma_v^2 + \phi_i$ 인 서로 독립인 정규분포를 따르게 된다. 이러한 분포에 대한 가정하에서 계산된 β 와 σ_v^2의 최대우도추정량을 제한최대우도추정량(REML : Restricted Maximum Likelihood Estimator)이라 하며, 선형혼합모형하에서도 근사적으로 유효하다. 따라서 $\hat{\theta}_i$ 의 BLUP 추정량을 산정 시 σ_v^2의 REML 추정량을 이용하여도 근사적으로 타당하다.

2.3.3 경험적 베이즈(EB) 방법

경험적 베이즈(EB) 추정법은 랜덤오차 v_i와 e_i가 정규분포를 따른다는 가정하에서 출발한다. $(\hat{\theta}_i, \theta_i)$의 결합분포는 평균이 $(z_i^T \beta, z_i^T \beta)$이고, 분산이 $(\sigma_v^2 + \phi_i, \sigma_v^2)$이며, 상관계수가 γ_i 인 이변량 정규분포를 따른다고 가정하자. 이때, θ_i 의 추정량의 평균제곱오차를 최소로하는 베이즈 추정량은 다음과 같다.

$$\hat{\theta}_i^B(\boldsymbol{\beta},\ \sigma_v^2) = E(\theta_i|\ \hat{\theta}_i,\ \boldsymbol{\beta},\ \sigma_v^2) = \gamma_i\hat{\theta}_i + (1-\gamma_i)\ \boldsymbol{z}_i^T\boldsymbol{\beta} \quad (2.25)$$

(2.25)식의 베이즈 추정량은 선형성 또는 불편성을 만족하지는 않는다. 여기에서 모수 $\boldsymbol{\beta}$ 와 σ_v^2을 제한최대우도(REML) 추정량으로 대체하여 다음과 같은 θ_i 의 경험적 베이즈(EB) 추정량을 얻는다.

$$\hat{\theta}_i^{EB} = \hat{\theta}_i^B(\hat{\boldsymbol{\beta}},\ \hat{\sigma}_v^2) \quad (2.26)$$

경험적 베이즈(EB) 추정량 $\hat{\theta}_i^{EB}$ 는 정규분포의 가정하에서는 EBLUP 추정량 $\hat{\theta}_i$ 와 같다. 그러나 경험적 베이즈방법은 $\hat{\theta}_i$ 과 θ_i 의 임의의 결합분포에 대해서도 일반적으로 응용할 수 있다는 점을 장점으로 들 수 있다.

EBLUP 추정량 $\hat{\theta}_i = \hat{\theta}_i(\hat{\sigma}_v^2)$의 *MSE* 추정량은 (2.22)식에서 σ_v^2 대신 $\hat{\sigma}_v^2$을 대체하여 얻어질 수 있으나, 이 경우에는 σ_v^2에 대한 추정효과가 무시되기 때문에 *MSE*의 추정값은 과소추정되는 경향을 보인다. 이러한 문제 때문에 Prasad and Rao(1990)는 $\{v_i\}$와 $\{e_i\}$에 대해 정규성을 가정하여 근사적으로 불편인 EBLUP 추정량 $\hat{\theta}_i$ 의 *MSE* 추정량을 제안하였다. Prasad and Rao(1990)가 제안한 *MSE* 추정량은 (2.23)식의 σ_v^2의 적률추정량을 사용하였을 경우 다음과 같이 주어진다.

$$mse(\hat{\theta}_i) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2) \quad (2.27)$$

단, $g_{1i}(\sigma_v^2) = \gamma_i\psi_i$, $g_{2i}(\sigma_v^2) = \sigma_v^2(1-\gamma_i)^2\ \boldsymbol{z}_i^T(\sum_i \gamma_i\ \boldsymbol{z}_i\ \boldsymbol{z}_i^T)^{-1}\boldsymbol{z}_i$,

$g_{3i}(\sigma_v^2) = \{\psi_i^2/(\sigma_v^2+\psi_i)^3\}h(\sigma_v^2)$, $h(\sigma_v^2) = 2m^{-2}\sum_i(\sigma_v^2+\psi_i)^2$

로 주어진다.

최근들어 Jing, Lahiri and Wan(1999)는 근사적으로 불편인 젝나이프(Jacknife) *MSE* 추정량을 제안하였다. 젝나이프 방법은 랜덤인 지역효과들을 갖는 로지스틱 회귀와 같은 좀 더 복잡한 모형들에 대해서도 쉽게 적용

할 수 있다는 장점을 갖고 있다.

θ_i 의 EB 추정량 (2.26)을 $\hat{\theta}_i^{EB} = k(\hat{\theta}_i, \hat{\varphi})$로 표현할 때, 젝나이프 절차는 다음과 같다. 여기에서 $\varphi = (\beta, \sigma_v^2)$은 모형에서의 모수 β 와 σ_v^2을 나타낸다.

(i) l 번째 지역의 자료 $(\hat{\theta}_l, z_l)$을 제외한 φ 의 추정량 $\hat{\varphi}(l)$을 계산한다. 이때의 EB 추정량을 $\hat{\theta}_i^{EB}(l) = k(\hat{\theta}_i, \hat{\varphi}(l))$로 나타내자.

(ii) $\hat{M}_{2i} = \frac{m-1}{m} \sum_{l=1}^{m} (\hat{\theta}_i^{EB}(l) - \hat{\theta}_i^{EB})$ 를 계산한다.

(iii) $\hat{M}_{1i} = g_{1i}(\hat{\sigma}_v^2) - \frac{m-1}{m} \sum_{l=1}^{m} \{g_{1i}(\hat{\sigma}_v^2(l) - g_{1i}(\hat{\sigma}_v^2)\}^2$ 을 계산한다.

(iv) 마지막으로 *MSE*의 젝나이프추정량 $mse_J(\hat{\theta}_i^{EB}) = \hat{M}_{1i} + \hat{M}_{2i}$를 계산한다.

$\hat{M}_{1i}$ 은 φ 가 기지일 때 *MSE*에 대한 추정량이며, $\hat{M}_{2i}$ 는 모형 모수 φ 를 추정할 때 추가적으로 발생하는 *MSE* 에 대한 변화량을 추정한다.

2.3.4 계층적 베이즈(HB) 방법

계층적 베이즈(HB : Hierarchical Bayes) 방법을 이용한 추론은 비교적 추론의 정확도가 높고, 복잡한 유형의 문제들에서도 최근에 개발된 MCMC(Markov Chain Monte Carlo)방법을 이용하여 해결할 수 있다. 깁스 샘플러도 이러한 방법의 일종이다. HB 방법에서는 모형 모수 $\varphi = (\beta, \sigma_v^2)$뿐만 아니라 모집단의 값 θ_i 가 랜덤으로 간주되며, 모형 모

수들에 대한 사전분포가 주어져야 한다. θ_i 들에 관한 추론은 주변 사후분포에 의해 결정된다. 즉, 주어진 자료 $\{(\hat{\theta}_i, z_i), i=1,2,\cdots,m\}$에 대한 조건부 분포 $f(\theta_i|\hat{\boldsymbol{\theta}})$에 의해 추론이 행해진다. 여기에서 $\hat{\boldsymbol{\theta}}$ 은 직접추정값 $\hat{\theta}_i$의 벡터이다. 특히 θ_i 는 사후분포의 평균 $E(\theta_i|\hat{\boldsymbol{\theta}})$에 의해 추정되며, 추정량의 변동은 사후분포의 분산 $V(\theta_i|\hat{\boldsymbol{\theta}})$에 의해 추정된다.

먼저 σ_v^2 이 기지인 상태를 가정하고 $\boldsymbol{\beta}$ 에 관한 사전분포를 배정하기로 한다. $\boldsymbol{\beta}$의 사전분포가 상수에 비례하고(i.e improper prior), v_i 와 e_i 가 정규분포를 따른다고 가정한다면, 이때 사후평균 $E(\theta_i|\hat{\boldsymbol{\theta}}, \sigma_v^2)$은 (2.20)식의 BLUP 추정량 $\tilde{\theta}_i(\sigma_v^2)$과 동일하다. 더욱이 사후분산 $V(\theta_i|\hat{\boldsymbol{\theta}}, \sigma_v^2)$은 (2.22)식의 BLUP 추정량의 MSE 와 같다. 따라서 σ_v^2이 기지인 상태에서는 HB 방법과 EBLUP 방법은 동일한 추론을 이끌어 낸다고 볼수 있다.

실제의 문제에서는 σ_v^2은 대부분 미지의 값으로 나타난다. 이러한 경우에는 $\boldsymbol{\beta}$ 뿐만 아니라 σ_v^2에 관한 사전분포를 고려해야 하며, 또한 서로 독립임을 가정하여 주변사후분포 $f(\sigma_v^2|\hat{\boldsymbol{\theta}})$을 이끌어 낸다. 만약 σ_v^2에 관한 사전분포로 불완전(improper) 사전분포를 배정한다면, θ_i 의 사후분포가 불완전 사후분포가 될 수 있기 때문에 이러한 문제를 피하기 위해서 $\tau_v = \sigma_v^{-2}$의 사전분포를 $G(a, b)$와 같이 배정한다(여기에서 $G(a, b)$는 감마함수로서 확률밀도함수는 $f(\tau_v) \propto \exp(-a\tau_v)\ \tau_v^{b-1}$의 형식이다). 주변사후확률분포 $f(\sigma_v^2|\hat{\boldsymbol{\theta}})$를 이용한 HB 추정량 $E(\theta_i|\hat{\boldsymbol{\theta}})$은 다음 식과 같이 주어진다.

$$\hat{\theta}_i^{HB} = E(\theta_i|\hat{\boldsymbol{\theta}}) = \int \tilde{\theta}_i(\sigma_v^2) f(\sigma_v^2|\hat{\boldsymbol{\theta}})\, d\sigma_v^2 \qquad (2.28)$$

위의 (2.28)식을 $E_{\sigma_v^2|\hat{\theta}}\{\tilde{\theta}_i(\sigma_v^2)\}$으로 표현하면, 사후분산 $V(\theta_i|\hat{\theta})$ 은 다음과 같다.

$$V(\theta_i|\hat{\theta}) = E_{\sigma_v^2|\hat{\theta}}\{g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2)\} + V_{\sigma_v^2|\hat{\theta}}\{\tilde{\theta}_i(\sigma_v^2)\} \quad (2.29)$$

여기에서 $V_{\sigma_v^2|\hat{\theta}}$ 은 $f(\sigma_v^2|\hat{\theta})$에 관한 분산을 의미한다.

위에서 소개한 (2.28)와 (2.29)은 일차원 수치적분에 의해 계산된다. 좀 더 복잡한 모형에 대한 고차원 수치적분은 MCMC 방법을 이용하여 계산할 수 있다. (2.28)식으로부터 $\tilde{\theta}_i^{HB}$ 는 EBLUP(EB) 추정량 $\tilde{\theta}_i(\hat{\sigma}_v^2)$와 근사적으로 같다는 것을 알 수 있다.

깁스 샘플링은 위의 식(2.28)와 (2.29)을 계산하는데 사용될 수 있는 일종의 MCMC 방법이다. 깁스 샘플링을 수행하기 위해서는 다음과 같은 깁스 조건부 분포들이 필요하다.

(i) $\beta|\theta,\ \sigma_v^2,\ \hat{\theta} \sim N_p((\sum z_i\ z_i^T)^{-1}(\sum z_i\theta_i),\ \sigma_v^2(\sum z_i\ z_i^T)^{-1})$

(ii) $\theta_i|\beta,\ \sigma_v^2,\ \hat{\theta} \sim N(\tilde{\theta}_i^B(\beta,\ \sigma_v^2),\ g_{1i}(\sigma_v^2)=\gamma_i\psi_i)$

(iii) $\tau_v = \sigma_v^{-2}|\beta,\ \theta,\ \hat{\theta} \sim G(\tilde{a},\ \tilde{b})$,

단, $\tilde{a}=\frac{1}{2}\sum(\theta_i - z_i^T)^2+a$, $\tilde{b}=\frac{m}{2}+b$.

깁스 샘플링 앨고리즘은 다음 절차에 의해 이루어진다.

(a) $\theta_i = \theta_i^{(0)}$, $\sigma_v^2 = \sigma_v^{2(0)}$을 초기값으로 하여 위의 (i)로부터 $\beta^{(1)}$을 발생시키고,

(b) $\beta = \beta^{(1)}$, $\sigma_v^2 = \sigma_v^{2(0)}$를 이용하여 위의 (ii)로부터 $\theta_i^{(1)}$ ($i=1,2,\cdots,m$) 을 생성하며,

(c) $\theta_i = \theta_i^{(1)}$과 $\beta = \beta^{(1)}$을 이용하여 위의 (iii)으로부터 $\sigma_v^{2(1)}$을 추출한다.

(d) 절차 (a), (b)와 (c)를 한 사이클로 하여 필요한 횟수만큼 반복 수행

생성된 확률변수값들이 안정적으로 되는 시점 t 까지 충분히 반복한 후, 이후부터 중심극한정리를 적용할 만큼 충분히 많은 크기인 J 개의 표본 $\{\beta^{(t+j)}, \sigma_v^{2(t+j)}, \theta_1^{(t+j)}, \cdots, \theta_m^{(t+j)}; j=1,2,\cdots,J\}$을 β, σ_v^2, θ_1, $\cdots$, θ_m의 결합 사후분포로 추출한 표본으로 간주한다. 초기값은 보통 $\theta_i^{(0)} = \hat{\theta}_i^{EB}$, $\sigma_v^{2(0)} = \sigma_v^2$의 REML 추정값을 사용한다.

위에서 계산된 J 개의 표본을 이용하여 θ_i 의 사후평균, 사후분산을 다음과 같이 산정한다.

$$\hat{\theta}_i^{HB} \approx \frac{1}{J}\sum_j \hat{\theta}_i(\sigma_v^{2(t+j)})$$

$$= \frac{1}{J}\sum_j \hat{\theta}_i(j) = \hat{\theta}_i(\cdot)\ , \qquad (2.30)$$

$$V(\theta_i|\hat{\theta}) \approx \frac{1}{J}\sum_j \{g_{1i}(\sigma_v^{2(t+j)}) + g_{2i}(\sigma_v^{2(t+j)})\} + \frac{1}{J}\sum_j \{\hat{\theta}_i(j) - \hat{\theta}_i(\cdot)\}^2 \qquad (2.31)$$

2.4 소지역 추정법 적용 사례

소지역 추정법이 다양한 분야에서 적용 가능함을 보이기 위해서 수치적인 적용 사례를 소개하고 이를 통해서 우리나라 시군구의 실업통계 작성법에 대한 가능성을 보이고자 한다.

2.4.1 다단수준 모형(Multi-level Model) 적용 사례

다단수준(Multilevel) 모형은 소지역 내의 변동과 소지역 간의 변동을 함께 고려하여 소지역 추정의 정확성을 높이기 위하여 제안되었던 방법이며, 조사자료를 이용하여 가정한 다단수준 모형의 모수를 추정하고 구하고자 하는 소지역의 특성값을 예측한다.

(1) 다단수준(Multi-level) 모형의 구조분석

다음과 같이 단위수준과 지역수준 공변량을 하나의 모형으로 통합한 다단수준 모형을 고려한다.

$$Y_i = X_i \beta_i + \varepsilon_i \ ,$$

$$\beta_i = Z_i \gamma + \nu_i \quad , \quad \text{단, } i = 1, 2, \cdots, m$$

여기에서 $Y_i = i$번째 소지역에서 길이가 n_i인 벡터, $X_i = i$번째 소지역에서 설명변수들의 행렬($(p+1) \times n_i$ 행렬), $Z_i =$ 소지역 변수들의 설계 행렬, $\gamma =$ 길이가 q인 벡터, $\beta_i = (\beta_{i0}, \beta_{i1}, \cdots, \beta_{ip})^t$는 i번째 소지역에서 길이가 $p+1$인 회귀계수의 벡터, $\nu_i = (\nu_{i0}, \nu_{i1}, \cdots, \nu_{ip})^t$는 i번째 소지역에서 길이가 $p+1$인 랜덤효과의 벡터를 나타내며, 오차항 $\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2 I)$,

$\nu_i \overset{ind}{\sim} N(0, \Omega)$ 를 가정하며, ε_i와 ν_i는 서로독립임을 가정한다.

i번째 소지역에 대해서 위의 모형을 구체적으로 행렬의 형식으로 표현하면 다음과 같이 나타낼 수 있다.

$$\begin{aligned} Y &= X\beta + \varepsilon \\ &= XZ_\gamma + X\nu + \varepsilon \ , \quad \text{단, } Z_\gamma = Z\gamma \end{aligned} \tag{2.32}$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ Y_i \\ \cdot \\ \cdot \\ Y_m \end{pmatrix} = \begin{pmatrix} X_1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & X_2 & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & X_i & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot & \cdot & X_m \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \beta_i \\ \cdot \\ \cdot \\ \beta_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \varepsilon_i \\ \cdot \\ \cdot \\ \varepsilon_m \end{pmatrix}$$

$$Y \qquad\qquad X \qquad\qquad \beta \qquad \varepsilon$$

여기에서 $Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \cdot \\ \cdot \\ \cdot \\ Y_{in_i} \end{pmatrix}$, $X_i = \begin{pmatrix} 1 & x_{i11} & x_{i21} & \cdot & \cdot & x_{ip1} \\ 1 & x_{i12} & x_{i22} & \cdot & \cdot & x_{ip2} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{i1n_i} & x_{i2n_i} & \cdot & \cdot & x_{ipn_i} \end{pmatrix}$

$$\beta_i = \begin{pmatrix} \beta_{i0} \\ \beta_{i1} \\ \cdot \\ \cdot \\ \cdot \\ \beta_{ip} \end{pmatrix} \ (= Z_i \gamma + \nu_i \ ) , \qquad \varepsilon_i = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{in_i} \end{pmatrix}$$

즉, 위의 모형은 회귀계수에 대한 랜덤효과를 모형에 포함시켜 단위수준과 지역수준 공변량을 한 모형으로 통합한 모형구조를 갖고 있다.

(2) 고정 모수 γ , 분산성분 모수 $\theta = ([\ \mathrm{Vech}\,(\Omega)]^t, \sigma^2)^t$ 의 추정

고정모수 γ 와 분산모수 θ 의 추정은 정규조건 하에서 ML 방법과 REML(Restricted ML) 방법을 이용하여 추정하는 방법, IGLS(Iterative Generalized Least Squares) 절차를 이용하여 추정하는 방법, RIGLS(Restricted IGLS) 절차를 이용하여 추정하는 방법이 있으며, 여기에서 IGLS 추정량은 일치추정량이 되고, 정규성 가정하에서 MLE 추정량과 동치이며, RIGLS 추정 절차는 분산성분 모수들의 불편추정량을 제공하며 정규성 가정하에서 REML 추정량과 동치이다.

IGLS 절차를 이용하여 γ 와 θ 의 추정하는 방법은 다음과 같다.

i) $\theta = ([\ \mathrm{Vech}(\Omega)]^t , \sigma^2)^t$ 의 초기값을 setting하여 $\tilde{\gamma}$ 를 다음식으로부터 산정한다.

$$\tilde{\gamma} = (Z^t X^t V^{-1} X Z)^{-1} (Z^t X^t V^{-1} Y)$$

$$= (\sum_{i=1}^{m} Z_i^t X_i^t V_i^{-1} X_i Z_i)^{-1} (\sum_{i=1}^{m} Z_i^t X_i^t V_i^{-1} Y_i) \qquad (2.33)$$

여기에서 $V_i = \sigma^2 I + X^t \Omega X$ 는 Y_i 의 공분산 행렬이며, $Z_i =$ design matrix, $V = Diag(V_1, V_2, \cdots, V_m) = Diag(V(Y_1), V(Y_2), \cdots, V(Y_m))$ 를 나타낸다.

ii) 위에서 추정된 $\tilde{\gamma}$ 값과 θ 의 초기값을 setting하여 θ 의 개선된 추정값 $\hat{\theta}_a$ 를 다음식으로 추정한다.

$$\hat{\theta}_a = Cov(\hat{\theta}_a)(\frac{\partial\ \mathrm{Vech}(V)}{\partial \theta})^t(\frac{1}{2} V^{-1} \otimes V^{-1})\ \mathrm{Vech}(\hat{Y}\ \hat{Y}^t), \qquad (2.34)$$

단, $\hat{Y} = Y - XZ_{\tilde{\gamma}}$, $Cov(\hat{\theta}_a) = \left\{ (\frac{\partial\ \mathrm{Vech}(V)}{\partial \theta})^t (\frac{1}{2} V^{-1} \otimes V^{-1}) (\frac{\partial\ \mathrm{Vech}(V)}{\partial \theta}) \right\}^{-1}$ 를 나타낸다.

모형 (2.32)를 가정하였을 때 γ 의 GLSE(Generalized Least Squares Estimator) $\tilde{\gamma}$ 는 (2.33)식과 같음을 보일 수 있으나, V 가 미지이므로 γ 는 (2.33)식을 이용하여 직접적으로 추정할 수는 없다. 따라서 θ 의 초기값을 대입하여 반복 알고리즘을 적용하면서 γ 를 추정한다.

추정된 $\tilde{\gamma}$ 를 이용하여 V 를 추정(θ 를 추정)함에 있어 다음의 사실을 이용하고, 먼저 γ 가 기지인 상태에서 V 를 추정한다.

i) $\mathrm{Vech}(V) = E[\ \mathrm{Vech}\{(Y - XZ_\gamma)(Y - XZ_\gamma)'\}]$

$= E[\ Y^*]$; Y^* 는 $\mathrm{Vech}(V)$의 불편추정량

ii) $\mathrm{Vech}(V)$는 $\theta=([\mathrm{Vech}(\Omega)]^t, \sigma^2)^t$ 의 선형함수

위의 i), ii)를 이용하여 다음과 같은 선형함수를 생각하자.

$$Y^* = F\theta + \zeta, \tag{2.35}$$

단, $F=\frac{\partial\,\mathrm{Vech}(V)}{\partial\theta}$, ζ는 확률변수로서 평균이 0이고, 분산이 $V_\zeta = 2\phi_n(V\otimes V)\phi_n{}^t$이다. 여기에서 ϕ_n 은 $\mathrm{Vech}(A)$에서 $\mathrm{Vech}(A)$로 가는 임의의 선형변환이고, A 는 $\mathrm{Vech}(A)=\phi_n\,\mathrm{Vech}(A)$ 를 만족하는 $n\times n$ 행렬이다.

위 (2.35)식의 Y^* 에서 V_ζ 를 기지이며, 정칙행렬로 가정하여 θ 의 GLSE를 (2.34)식과 같이 추정한다. 이때 (2.34)식의 $\tilde{\theta}_a$ 는 θ 의 초기값을 대입하여 반복적으로 계산 추정된다.

RIGLS 절차를 이용하여 γ 와 θ 를 추정하는 방법은 다음과 같다.

i) $\theta=([\mathrm{Vech}(\Omega)]^t, \sigma^2)^t$ 의 초기값을 setting하여 $\tilde{\gamma}$ ($=\hat{\gamma}$)를 산정한다(IGLS 절차에서 $\tilde{\gamma}$ 추정과 동일함).

ii) i)에서 추정된 $\hat{\gamma}$ 값과, θ 의 초기값을 setting하여 $\tilde{\theta}_a$ 를 계산하는데, 여기에서는 IGLS의 (2.34)식의 항들 중,

$$\tilde{Y}\,\tilde{Y}^t = (Y - XZ_{\hat{\gamma}})(Y - XZ_{\hat{\gamma}})^t$$

대신에

$$\tilde{Y}\,\tilde{Y}^t = (Y - XZ_{\hat{\gamma}})(Y - XZ_{\hat{\gamma}})^t + XZ(Z^tX^tV^{-1}XZ)^{-1}Z^tX^t$$

를 이용한다.

V 가 기지인 상태에서 GLS방법으로 γ 가 추정된다면 다음의 식

$E\{(Y - XZ_{\hat{\gamma}})(Y - XZ_{\hat{\gamma}})^t\} = V - XZ(Z^tX^tV^{-1}XZ)^{-1}Z^tX^t$ 이 성립한

다. V 의 근사적인 불편추정량 (즉, θ 의 불편추정량)을 얻기위한 방법으로 반복계산에서 다음식을 이용한다.

$$\hat{Y}\ \hat{Y}^t = (Y - XZ\hat{\gamma})(Y - XZ\hat{\gamma})^t + XZ(Z^tX^tV^{-1}XZ)^{-1}Z^tX^t$$

(3) 소지역 평균의 추정량

① μ_i의 EBLUP 추정량

소지역의 특성치를 추정하기 위하여 먼저 다음과 같은 모형을 가정하기로 한다.

$$Y_i = X_i\beta_i + \varepsilon_i\ ,$$

$$\beta_i = Z_i\gamma + \nu_i \quad , \quad i = 1, 2, \cdots, m \text{ (소지역 수)}$$

여기에서 N_i는 i 번째 소지역의 모집단 크기를 나타낸다.

위의 모형에서 i 번째 소지역에 대한 평균의 기댓값은 다음의 (2.36)식과 같이 표현할 수 있다.

$$\begin{aligned}\mu_i &= \overline{X}_i^t\beta_i \\ &= \overline{X}_i^t(Z_i\gamma + \nu_i) \\ &= \overline{X}_i^tZ_i\gamma + \overline{X}_i^t\nu_i\ ,\end{aligned} \tag{2.36}$$

여기에서 $\overline{X}_i = (1, \bar{x}_{i1}, \bar{x}_{i2}, \cdots, \bar{x}_{ip})^t$ 는 i 번째 소지역에서 설명변수의 모집단 평균벡터를 나타낸다.

μ_i 의 EBLUP 추정량 $\hat{\mu}_i$ 은 다음의 (2.37)식과 같이 표현된다.

$$\hat{\mu}_i = \overline{X}_i^tZ_i\hat{\gamma} + \overline{X}_i^t\hat{\nu}_i\ , \tag{2.37}$$

단, $\hat{\nu}_i = \hat{\Omega}\ X_i^t\ \hat{V}_i^{-1}(Y_i - X_iZ_i\hat{\gamma})$ $\hat{V}_i^{-1} = \hat{\sigma}^{-2}I - \hat{\sigma}^{-4}X_i\hat{\Omega}\ \hat{G}_i^{-1}\ X_i^t$,

$\hat{G}_i^{-1} = (I + \hat{\sigma}^{-2}\ X_i^tX_i\hat{\Omega})^{-1}$이며, 여기에서 γ 와 θ 의 추정량 $\hat{\gamma}$ 와

$\hat{\theta}$ 는 앞서 계산된 RIGLS 추정값을 이용한다.

절편항만 랜덤이고, β 의 나머지 항들은 고정계수인 모형을 Battese et al.(1981, 1988)이 제안하였고 이모형에서 μ_i 의 EBLUP 추정량 $\hat{\mu}_{i(RI)}$ 은 다음의 (2.38)식과 같다.

$$\hat{\mu}_{i(RI)} = \overline{X}_i^t \hat{\beta} + \hat{\nu}_{i0} \tag{2.38}$$

② $\hat{\mu}_i$의 *MSE* 근사와 추정식

다음의 사실을 이용하여 $\hat{\mu}_i$ 의 MSE 근사식을 유도한다.

i) μ_i 의 REMLE 는 translation invariant (Kackar & Harville(1984)).

ii) μ_i 의 RIGLSE 는 정규성 가정 하에서 REMLE와 동치 (Goldstein(1989))

위의 사실을 이용한다면, μ_i 의 RIGLSE 인 $\hat{\mu}_i$ 도 translation invariant 이므로 다음 식이 성립하게 된다.

$$\begin{aligned} MSE(\hat{\mu}_i) &= E(\hat{\mu}_i - \mu_i)^2 \\ &= E(\tilde{\mu}_i - \mu_i)^2 + E(\hat{\mu}_i - \tilde{\mu}_i)^2 , \quad i = 1, 2, \cdots, m \end{aligned}$$

단, $\tilde{\mu}_i$ 는 μ_i 의 BLUP 추정량이며, 여기에서 첫 번째 항인 $E(\tilde{\mu}_i - \mu_i)^2$은 다음 (2.39)식의 결과에서 계산된다.

$$\begin{aligned} MSE(\tilde{\mu}_i) &= E(\tilde{\mu}_i - \mu_i)^2 \\ &= \overline{X}_i^t (G_i^{-1})^t \Omega \overline{X}_i \\ &\quad + \sigma^2 \overline{X}_i (G_i^{-1})^t Z_i \left(\sum_{i=1}^{m} Z_i^t G_i^{-1} X_i^t X_i Z_i \right)^{-1} Z_i^t G_i^{-1} \overline{X}_i \\ &= T_1 + T_2 , \end{aligned} \tag{2.39}$$

단, $G_i = I + \sigma^2 X_i^t X_i \Omega$

두 번째 항 $E(\widehat{\mu_i}-\widetilde{\mu_i})^2$은 (2.40)식과 같이 근사적인 계산식으로 주어진다.

$$E(\widehat{\mu_i}-\widetilde{\mu_i})^2 \approx T_3$$

$$= tr\Big[\Big(\frac{\partial d_i}{\partial \theta}\Big) V \Big(\frac{\partial d_i}{\partial \theta}\Big)^t E\{(\hat{\theta}-\theta)(\hat{\theta}-\theta)^t\}\Big] \qquad (2.40)$$

여기에서 $d_i = \overline{X_i}^t K_i (I \otimes \Omega) X^t V^{-1}$, $K_i = [0, \cdots, 0, I, 0, \cdots, 0]$는 $(p+1)\times(p+1)m$ 행렬로써 i 번째 소지역에서 $(p+1)\times(p+1)$ 단위행렬을 가지며, $\hat{\theta} = (\hat{\theta_1}, \hat{\theta_2}, \cdots, \hat{\theta_s})$은 $\theta = (\theta_1, \theta_2, \cdots, \theta_s)$의 translation invariant 추정량으로써, $\theta_s = \sigma^2$, $\theta_k (k=1,2,\cdots,s-1)$는 Ω의 서로다른 원소를 나타낸다.

위의 (2.40)식에서 $E\{(\hat{\theta}-\theta)(\hat{\theta}-\theta)^t\}$ 를 REML 추정량 B의 점근적 공분산 행렬로 근사시킨 계산 결과식은 다음 (2.40)식과 같다.

$$T_3 = \overline{X_i}^t (G_i^{-1}) \Big(\sum_{j=1}^{s-1}\sum_{k=1}^{s-1} b_{jk} \triangle_j C_i \triangle_k^t\Big) G_i^{-1} \overline{X_i}$$

$$- 2\,\overline{X_i}^t (G_i^{-1})^t \Big(\sum_{j=1}^{s-1} b_{j,s} \triangle_j\Big) R_i \Omega \overline{X_i} + b_{ss}\, \overline{X_i}^t \Omega S_i \Omega \overline{X_i}\,, \qquad (2.41)$$

여기에서 $B = \theta$ 의 REML 추정량, $b_{j,k} = B$ 의 jk 번째 원소, $b^*_{jk} = B^{-1}$ 의 jk 번째 원소 $= tr\Big(\sum_{i=1}^{m} P_i \frac{\partial V}{\partial \theta_j} P_i \frac{\partial V}{\partial \theta_k}\Big)$ $(k=1,2,\cdots,s)$ 이며, P_i는 다음과 같이 표현된다.

$$P_i = V_i^{-1} - V_i^{-1} X_i Z_i \Big(\sum_{i=1}^{m} Z_i^t X_i^t V_i^{-1} X_i Z_i\Big) Z_i^t X_i^t V_i^{-1},$$

단, $C_i = \sigma^{-2} G_i^{-1} X_i^t X_i$,

$R_i = \sigma^{-4} G_i^{-2} X_i^t X_i$,

$S_i = \sigma^{-6} G_i^{-3} X_i^t X_i$,

$$\triangle_k = \frac{\partial \Omega}{\partial \Omega_k} \ : \ (s-1)\times(s-1) \ \ (k=1,2,\cdots,s-1).$$

위의 (2.39)식과 (2.41)식을 정리하면, $\hat{\mu}_i$ 의 *MSE* 근사식은 다음 (2.42)식과 같이 표현된다.

$$\begin{aligned} MSE(\hat{\mu}_i) &= E(\hat{\mu}_i - \mu_i)^2 \\ &= E(\tilde{\mu}_i - \mu_i)^2 + E(\hat{\mu}_i - \tilde{\mu}_i)^2 \ , \quad i=1,2,\cdots,m \\ &\approx \left\{ \overline{X}_i^t (G_i^{-1})^t \Omega \overline{X}_i \right\} \\ &+ \left\{ \sigma^2 \overline{X}_i (G_i^{-1})^t Z_i \left(\sum_{i=1}^{m} Z_i^t G_i^{-1} X_i^t X_i Z_i \right)^{-1} Z_i^t G_i^{-1} \overline{X}_i \right\} \\ &+ \left\{ \overline{X}_i^t (G_i^{-1}) \left(\sum_{j=1}^{s-1} \sum_{k=1}^{s-1} b_{jk} \triangle_j C_i \triangle_k^t \right) G_i^{-1} \overline{X}_i \right. \\ &\left. - 2 \overline{X}_i^t (G_i^{-1})^t \left(\sum_{j=1}^{s-1} b_{j,s} \triangle_j \right) R_i \Omega \overline{X}_i + b_{ss} \overline{X}_i^t \Omega S_i \Omega \overline{X}_i \right\} \\ &= T_1 + T_2 + T_3 \end{aligned} \tag{2.42}$$

한편, 위의 결과들을 유한 모집단에 적용할 경우에는 다음과 같이 보정 절차를 거쳐 표본으로 추출되지 않은 단위들의 영향력을 고려해 주어야 한다.

$$\hat{\mu}_i^F = f_i \bar{y}_i + (\overline{X}_i - f_i \bar{x}_i)^t (Z_i \hat{\gamma} + \hat{\nu}_i) \ , \tag{2.43}$$

단, $f_i = n_i / N_i$, $\bar{x}_i = (1, \bar{x}_{i1}, \bar{x}_{i2}, \cdots, \bar{x}_{ip})^t$ 는 표본평균을 나타낸다.

$$\hat{\mu}_i^F - \overline{Y} = (1-f_i)\left[(\overline{X}_i^C)^t \left\{ Z_i(\hat{\gamma}-\gamma) + \hat{\nu}_i - \nu_i - \bar{\varepsilon}_i^C \right\} \right] , \tag{2.44}$$

여기에서 $\overline{X}_i^C = (1-f_i)^{-1}(\overline{X}_i - f_i \bar{x}_i)$, $\bar{\varepsilon}_i^C$ = 'i 번째 소지역에서 표본으로 추출되지 않은 단위들에 대한 ε_{ij} 의 평균'을 나타낸다.

위의 (2.43)식과 (2.44)식으로부터 추정량의 *MSE*를 계산하면 다음 (2.45)식과 같이 주어진다.

$$MSE(\hat{\mu}_i^F) = (1-f_i)^2 \{ MSE^*(\hat{\mu}_i) + N_i^{-1}(1-f_i)^{-1}\sigma^2 \}, \tag{2.45}$$

여기에서 $MSE^*(\hat{\mu}_i)$는 (2.42)식에서 $\overline{X}_i$ 를 $\overline{X}_i^C$ 로 대체하여 구한다.

$\hat{\mu}_i$ 의 MSE 추정식은 $\hat{\mu}_i$ 의 MSE 근사식 (2.42)과 (2.45)식에서 σ^2과 Ω 의 RIGLS 추정량을 이용할 수 있으며, RIGLS 추정량을 이용한 $\hat{\mu}_i$ 의 MSE 추정식은 (2.46)와 (2.47)과 같이 주어진다.

$$\widehat{MSE}(\hat{\mu}_i) = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 , \tag{2.46}$$

$$\widehat{MSE}(\hat{\mu}_i^F) = (1-f_i)^2 \{ \widehat{MSE}^*(\hat{\mu}_i) + N_i^{-1}(1-f_i)^{-1}\hat{\sigma}^2 \} \tag{2.47}$$

(4) 소지역 추정에 적용

Moura et al.(1999)는 다단수준 모형이 소지역 추정에 광범위하게 적용될 수 있는 가능성을 최근 논문에 발표하였다. 브라질의 하나의 county내의 전체 지역에서 조사된 자료를 이용하여 다단수준 모형하에서 미지인 모수들을 앞서 소개하였던 방법에 의해 추정하고, 전체 자료와 전체 자료 중 일부분을 랜덤 추출하여 얻은 표본 자료를 이용하여 소지역의 평균에 관한 추정치의 정확도를 비교하여 소지역 추정에 다단수준 모형이 적용될 수 있다는 가능성을 확인해 주었다.

사용된 자료는 브라질의 한 County내의 전체 지역에서 조사된 38,740 가구에 대한 조사자료이며, 종속변수는 가장의 수입액, 설명변수는 가장의 교육정도(0 ~ 5)와 가구별 방의 수(1 ~ 11)이다.

① 모수 추정

소지역 특성치를 추정하기 위해 사용된 다단수준 모형은 다음과 같다.

$$Y_{ij} = \beta_{i0} + \beta_{i1}x_{1ij} + \beta_{i2}x_{2ij} + \epsilon_{ij} , \tag{2.48}$$

$$\beta_{i0} = \gamma_{00} + \nu_{i0} , \ \beta_{i1} = \gamma_{10} + \nu_{i1} , \ \beta_{i2} = \gamma_{20} + \nu_{i2} ,$$

단, $i = 1, 2, \cdots, m$(소지역 수), $j = 1, 2, \cdots, N_i$ (i 번째 소지역의 전체

조사 가구수), $\varepsilon_{ij} \overset{ind}{\sim} N(0, \sigma^2)$, $\nu_i = (\nu_{i0}, \nu_{i1}, \nu_{i2}) \overset{ind}{\sim} N(0, \Omega)$

x_{1ij} 는 방의 수, x_{2ij} 는 가장의 교육정도를 나타내며 오차항 ε_{ij}와 ν_i는 서로 독립임을 가정한다. 여기에서 x_{1ij} 와 x_{2ij} 는 각각의 모평균에 대해서 표준화된 값을 나타낸다.

county내의 전체자료를 이용하여 앞서 소개되었던 방법에 근거하여 고정 모수 γ 와 분산성분 모수 $\theta = ([\text{Vech}(\Omega)]^t, \sigma^2)^t$ 의 RIGLS 추정량을 계산하면 결과는 다음과 같다.

i) 고정 모수 γ

$\gamma_{00} = 8.456$ (표준오차 : 0.108), $\gamma_{10} = 1.223$ (0.046) ,

$\gamma_{20} = 2.596$ (0.086)

ii) 분산성분 모수 θ

$\sigma^2 = 47.74$ (0.345) ,

$$V(\nu_i) = \begin{pmatrix} \sigma_{00} & \sigma_{01} & \sigma_{02} \\ \sigma_{10} & \sigma_{11} & \sigma_{12} \\ \sigma_{20} & \sigma_{21} & \sigma_{22} \end{pmatrix}$$

$$= \begin{pmatrix} 1.385(0.194) & 0.345(0.66) & 0.492(0.117) \\ & 0.234(0.35) & 0.333(0.054) \\ & & 0.926(0.124) \end{pmatrix}$$

(2) 모의실험을 통한 모형 적용 검토

수치조사를 시행하기 위하여 먼저 다음과 같은 자료생성 모형을 정의하기로 한다.

i) 일반 모형(G)

(5.17) 모형에서 $\nu_i = (\nu_{i0}, \nu_{i1}, \nu_{i2}) \overset{ind}{\sim} N(0, \Omega)$ 이고,

$$\Omega = \begin{pmatrix} 1.385(0.194) & 0.345(0.66) & 0.492(0.117) \\ & 0.234(0.35) & 0.333(0.054) \\ & & 0.926(0.124) \end{pmatrix}$$ 인 모형

ii) 대각 모형(D)

(5.17) 모형에서 $\nu_i = (\nu_{i0}, \nu_{i1}, \nu_{i2}) \overset{ind}{\sim} N(0, \Omega)$ 이고,

$$\Omega = \begin{pmatrix} 1.385(0.194) & 0 & 0 \\ 0 & 0.234(0.35) & 0 \\ 0 & 0 & 0.926(0.124) \end{pmatrix}$$ 인 모형

iii) 랜덤 절편항 모형(RI)

(5.17) 모형에서 $\nu_i = (\nu_{i0}, \nu_{i1}, \nu_{i2}) \overset{ind}{\sim} N(0, \Omega)$ 이고,

$$\Omega = \begin{pmatrix} 1.385(0.194) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$ 인 모형

가정된 자료생성 모형 m_1(= G, D, RI 모형)의 각각에 대해서 각 소지역별 평균의 기대값을 다음 식을 이용하여 반복 생성하고,

$$\mu_{im_1}^{(r)} = \beta_{0i}^{(r)} + \beta_{1i}^{(r)} \overline{X}_{1i} + \beta_{2i}^{(r)} \overline{X}_{2i},$$

단, $r = 1, 2, \cdots, R(=5000)$,

$\overline{X}_{1i}$, $\overline{X}_{2i}$: 각 소지역 전체 조사자료에 대한 평균

비교를 위하여 각 소지역의 전체 조사자료에 대해서 10%의 Simulation Subset을 추출하여 위에서 정의된 각 모형별로 Y_{ij} 를 생성하고, 모형을 적합시켜 각 소지역별 평균의 기대값에 대한 추정치를 다음 식을 이용하여 반복 생성하여,

$$\hat{\mu}_{im_1}^{(r)} = \hat{\beta}_{0i}^{(r)} + \hat{\beta}_{1i}^{(r)} \bar{x}_{1i} + \hat{\beta}_{2i}^{(r)} \bar{x}_{2i} ,$$

단, $r = 1, 2, \cdots, R(=5000)$,

$\bar{x}_{1i}$, $\bar{x}_{2i}$: 각 소지역의 Simulation Subset자료에 대한 평균

$\hat{\mu}_{im_1}^{(r)}$ 과 $\mu_{im_1}^{(r)}$ 의 차에 대한 양을 비교하고자 한다.

자료생성 모형 m_1(G, D, RI 모형)에 의해 생성된 자료를 이용하여 적합시킨 추정모형을 m_2(G, D, RI 추정모형)라 했을 때, 다음의 측도를 정의하자.

$$MSE(\hat{\mu}_{im_1, m_2}) = \frac{\sum_{r=1}^{R} (\hat{\mu}_{im_1, m_2}^{(r)} - \mu_{im_1}^{(r)})^2}{R} ,$$

$$ARE(\hat{\mu}_{im_1, m_2}) = \frac{\sum_{r=1}^{R} |\hat{\mu}_{im_1, m_2}^{(r)} - \mu_{im_1}^{(r)}| / \mu_{im_1}^{(r)}}{R} ,$$

$$\text{RMSE}_{m_2, m_1} = \sum_{i=1}^{m} \frac{MSE(\hat{\mu}_{im_1, m_2})}{MSE(\hat{\mu}_{im_1, m_1})} \times 100 ,$$

$$RARE_{m_2, m_1} = \sum_{i=1}^{m} \frac{ARE(\hat{\mu}_{im_1, m_2})}{ARE(\hat{\mu}_{im_1, m_1})} \times 100 .$$

여기에서 $MSE(\hat{\mu}_{im_1, m_2})$와 $ARE(\hat{\mu}_{im_1, m_2})$는 전체자료 중 10%의 Simulation Subset자료를 이용하여 m_1 모형에서 자료를 생성하여 m_2 모형을 적합시켰을 때 i소지역의 평균에 관한 추정치의 평균제곱오차 및 절대상대오차를 각각 의미하며, RMSE_{m_2, m_1}과 $RARE_{m_2, m_1}$는 m_1 모형에서 자료를 생성하여 m_1 모형을 추정했을때와 m_1 모형에서 자료를 생성하여 m_2 모형을 추정했을때의 추정량의 값에 대한 평균제곱오차비, 절대허용오차비를 각각 의미한다.

모의실험의 절차를 단계적으로 살펴보면 다음과 같이 설명할 수 있다.

제 1단계

Step1

일반모형(G)에서 명시된 $\nu_i = (\nu_{i0}, \nu_{i1}, \nu_{i2})$ 의 분포로부터 $\nu_1, \nu_2, \cdots, \nu_m$ 을 생성한다.

Step2

Step1에서 생성된 ν_i ($i = 1, 2, \cdots, m$) 에 대해서 $\beta_1, \beta_2, \cdots, \beta_m$ 을 결정한다. 여기에서 $\beta_i = (\beta_{i0}, \beta_{i1}, \beta_{i2})$, $\beta_{i0} = \gamma_{00} + \nu_{i0}$, $\beta_{i1} = \gamma_{10} + \nu_{i1}$, $\beta_{i2} = \gamma_{20} + \nu_{i2}$ 이며, 고정모수 $\gamma_{00} = 8.456$, $\gamma_{10} = 1.223$, $\gamma_{20} = 2.596$는 이미 주어진 값이다.

Step3

이미 계산된 i 번째 소지역의 전체 조사자료에 대한 평균들 ($\overline{X}_{1i}$, $\overline{X}_{2i}$) 와 Step2에서 결정된 $\beta_i = (\beta_{i0}, \beta_{i1}, \beta_{i2})$를 사용하여 일반모형(G)에 대한 i 번째 소지역의 평균에 대한 기대값 $\mu_{iG}{}^{(1)}$을 생성한다.

$$\mu_{iG}{}^{(1)} = \beta_{0i} + \beta_{1i}\,\overline{X}_{1i} + \beta_{2i}\,\overline{X}_{2i}\,,$$

여기에서 $i = 1, 2, \cdots, m$(소지역 수) , G 는 일반모형을 나타낸다. 즉 일반모형에 대해서 $\mu_{1G}{}^{(1)}$, $\mu_{2G}{}^{(1)}$, $\cdots$, $\mu_{mG}{}^{(1)}$이 생성된다.

Step4

Step1 ~ Step3 의 과정을 5000번 반복하여 General Model 의 i 번째 소지역의 평균에 대한 기대값 μ_{iG} 를 다음 식을 이용하여 5000개씩 생성한다.

$$\mu_{iG}{}^{(r)} = \beta_{0i}{}^{(r)} + \beta_{1i}{}^{(r)}\,\overline{X}_{1i} + \beta_{2i}{}^{(r)}\,\overline{X}_{2i}\,,$$

단, $i = 1, 2, \cdots, m$ (소지역 수), $r = 1, 2, \cdots, R(= 5000)$

Step5

대각모형(D), 랜덤절편항 모형(RI)에 대해서도 위의 Step1~Step4를 5000번 반복하여 $\mu_{iD}^{(r)}$, $\mu_{iRI}^{(r)}$을 생성한다.

제 2단계

Step1

제 1단계의 Step1 ~ Step2에서 얻어진 $\beta_1, \beta_2, \cdots, \beta_m$ 을 취한다.

Step2

각각의 소지역으로부터 10%의 표본을 랜덤 추출(=Simulation Subset)하여 보조 정보 (x_{1ij}, x_{2ij})를 확보하고, $\varepsilon_{ij} \sim N(0, \sigma^2=47.74)$로부터 보조 정보의 개수만큼 ε_{ij} 를 생성한다. 여기에서 $j(=1,2,\cdots,n_i)$는 i 번째 소지역에서 추출된 Simulation Subset의 개수를 나타낸다.

Step3

Step1~Step2에서 생성된 β_i , (x_{1ij}, x_{2ij}) , ε_{ij} 값을 다음 식에 대입하여 n_i 개의 Y_{ij} 값을 생성한다.

$$Y_{ij} = \beta_{i0} + \beta_{i1}x_{1ij} + \beta_{i2}x_{2ij} + \varepsilon_{ij}$$

Step4

i 번째 소지역에서 랜덤 추출된 n_i 개의 (x_{1ij}, x_{2ij}) 값과 이에 대응되는 n_i 개의 Y_{ij} 값을 이용하여 다음과 같은 세가지 모형(일반모형(G), 대각모형(D), 랜덤절편항 모형(RI))을 추정함

o 일반모형(G)에서 자료를 생성하여 일반모형(G)을 추정

$$\hat{Y}_{iG,G}^{(1)} = \hat{\beta}_{i0}^{(1)} + \hat{\beta}_{i1}^{(1)}x_{1i} + \hat{\beta}_{i2}^{(1)}x_{2i}$$

o 일반모형(G)에서 자료를 생성하여 대각모형(D)을 추정

$$\hat{Y}_{iG,D}^{(1)} = \hat{\beta}^{*}_{i0}{}^{(1)} \hat{\beta}^{*}_{i1}{}^{(1)}x_{1i} + \hat{\beta}^{*}_{i2}{}^{(1)}x_{2i}$$

o 일반모형(G)에서 자료를 생성하여 랜덤절편항 모형(RI)을 추정

$$\hat{Y}_{iG,RI}^{(1)} = \hat{\beta}^{**}_{i0}{}^{(1)} \hat{\beta}^{**}_{i1}{}^{(1)} x_{1i} + \hat{\beta}^{**}_{i2}{}^{(1)} x_{2i}$$

Step5

Step4 에서 계산된 회귀계수의 추정치를 이용하여 세가지 모형에 대해서 i 번째 소지역의 평균에 대한 기대값의 추정치를 다음 식으로 계산한다.

o 일반모형(G)에서 자료를 생성하여 일반모형(G)을 추정한 경우

$$\hat{\mu}_{iG,G}^{(1)} = \hat{\beta}_{i0}^{(1)} + \hat{\beta}_{i1}^{(1)} \bar{x}_{1i} + \hat{\beta}_{i2}^{(1)} \bar{x}_{2i}$$

o 일반모형(G)에서 자료를 생성하여 대각모형(D)을 추정한 경우

$$\hat{\mu}_{iG,D}^{(1)} = \hat{\beta}^{*}_{i0}{}^{(1)} \hat{\beta}^{*}_{i1}{}^{(1)} \bar{x}_{1i} + \hat{\beta}^{*}_{i2}{}^{(1)} \bar{x}_{2i}$$

o 일반모형(G)에서 자료를 생성하여 랜덤절편항 모형(RI)을 추정한 경우

$$\hat{\mu}_{iG,RI}^{(1)} = \hat{\beta}^{**}_{i0}{}^{(1)} \hat{\beta}^{**}_{i1}{}^{(1)} \bar{x}_{1i} + \hat{\beta}^{**}_{i2}{}^{(1)} \bar{x}_{2i}$$

여기에서 $\bar{x}_{1i}$, $\bar{x}_{2i}$ 는 Simulation Subset으로 추출된 x_{1ij} , x_{2ij} 의 평균을 나타낸다.

Step6

Step1 ~ Step5 의 절차를 5000번 반복하여 $\hat{\mu}_{iG,G}^{(r)}$, $\hat{\mu}_{iG,D}^{(r)}$, $\hat{\mu}_{iG,RI}^{(r)}$를 생성한다.

Step7

o 대각모형(D)을 가정하여 Step1의 절차에 따라 β_1 , β_2 , … , β_m 을 취하고, Step2 ~Step6의 절차를 수행하여 $\hat{\mu}_{iD,G}^{(r)}$, $\hat{\mu}_{iD,D}^{(r)}$, $\hat{\mu}_{iD,RI}^{(r)}$를 생성한다.

o 랜덤절편항 모형(RI)을 가정하여 Step1의 절차에 따라

$\beta_1, \beta_2, \cdots, \beta_m$ 을 취하고, Step2~Step6의 절차를 수행하여 $\hat{\mu}_{iRI,G}^{(r)}$, $\hat{\mu}_{iRI,D}^{(r)}$, $\hat{\mu}_{iRI,RI}^{(r)}$를 생성한다.

제 3단계

o $MSE(\hat{\mu}_{im_1,m_2}) = \dfrac{\sum_{r=1}^{R}(\hat{\mu}_{im_1,m_2}^{(r)} - \mu_{im_1}^{(r)})^2}{R}$ 계산,

; i.e $MSE(\hat{\mu}_{iG,G})$, $MSE(\hat{\mu}_{iG,D})$, $MSE(\hat{\mu}_{iG,RI})$,

$MSE(\hat{\mu}_{iD,G})$, $MSE(\hat{\mu}_{iD,D})$, $MSE(\hat{\mu}_{iD,RI})$,

$MSE(\hat{\mu}_{iRI,G})$, $MSE(\hat{\mu}_{iRI,D})$, $MSE(\hat{\mu}_{iRI,RI})$ 를 계산

$ARE(\hat{\mu}_{im_1,m_2}) = \dfrac{\sum_{r=1}^{R}|\hat{\mu}_{im_1,m_2}^{(r)} - \mu_{im_1}^{(r)}| / \mu_{im_1}^{(r)}}{R}$ 계산

; i.e $ARE(\hat{\mu}_{iG,G})$, $ARE(\hat{\mu}_{iG,D})$, $ARE(\hat{\mu}_{iG,RI})$,

$ARE(\hat{\mu}_{iD,G})$, $ARE(\hat{\mu}_{iD,D})$, $ARE(\hat{\mu}_{iD,RI})$,

$ARE(\hat{\mu}_{iRI,G})$, $ARE(\hat{\mu}_{iRI,D})$, $ARE(\hat{\mu}_{iRI,RI})$ 를 계산

o $\mathrm{RMSE}_{m_2,m_1} = \sum_{i=1}^{m} \dfrac{MSE(\hat{\mu}_{im_1,m_2})}{MSE(\hat{\mu}_{im_1,m_1})} \times 100$ 을 계산,

; i.e $\mathrm{RMSE}_{G,G}$, $\mathrm{RMSE}_{G,D}$, $\mathrm{RMSE}_{G,RI}$,

$\mathrm{RMSE}_{D,G}$, $\mathrm{RMSE}_{D,D}$, $\mathrm{RMSE}_{D,RI}$,

$\mathrm{RMSE}_{RI,G}$, $\mathrm{RMSE}_{RI,D}$, $\mathrm{RMSE}_{RI,RI}$ 를 계산

$RARE_{m_2,m_1} = \sum_{i=1}^{m} \dfrac{ARE(\hat{\mu}_{im_1,m_2})}{ARE(\hat{\mu}_{im_1,m_1})} \times 100$ 을 계산

; i.e $RARE_{G,G}$, $RARE_{G,D}$, $RARE_{G,RI}$,

$RARE_{D,G}$, $RARE_{D,D}$, $RARE_{D,RI}$,

$RARE_{RI,G}$, $RARE_{RI,D}$, $RARE_{RI,RI}$ 를 계산

모의실험 결과를 요약하면 다음의 <표 2.1>과 같다.

<표 2.1> 자료생성모형 $m_1 = G, D, RI$ 와 추정모형

$m_2 = G, D, RI$ 에 대한 RMSE ($RARE$) 값의 비교

Estimator	Data Generation Model		
	일반모형(G)	대각모형(D)	랜덤절편항모형 (RI)
General (G)	100.0 * (100.0)	101.8 ** (100.9)	101.2 (100.6)
Diagonal(D)	108.8 (82.6)	100.0 (100.0)	100.2 (100.1)
Random Intercept (RI)	131.9 (176.9)	109.1 (105.6)	100.0 (100.0)

$$(*)\quad \mathrm{RMSE}_{G,G} = \sum_{i=1}^{m} \frac{MSE(\hat{\mu}_{iG,G})}{MSE(\hat{\mu}_{iG,G})} \times 100 = 100.0$$

$$(**)\quad \mathrm{RMSE}_{G,D} = \sum_{i=1}^{m} \frac{MSE(\hat{\mu}_{iD,G})}{MSE(\hat{\mu}_{iD,D})} \times 100$$

$$= \frac{\sum_{i=1}^{m}\left\{\frac{\sum_{r=1}^{5000}(\hat{\mu}_{iD,G}^{(r)} - \mu_{iD}^{(r)})^2}{R}\right\}}{\sum_{i=1}^{m}\left\{\frac{\sum_{r=1}^{5000}(\hat{\mu}_{iD,D}^{(r)} - \mu_{iD}^{(r)})^2}{R}\right\}} \times 100 = 101.8$$

랜덤절편항 모형(RI)과 같은 단순한 모형에서 자료를 생성하여, 이 모형보다 복잡한 일반모형(G), 대각모형(D)을 적합시켜 얻은 추정량들의 평균제곱오차비(RMSE)를 살펴보면, $\mathrm{RMSE}_{G,RI} = 101.2$, $\mathrm{RMSE}_{D,RI} = 100.2$로써 심

한 효율의 손실은 발생하지 않았음을 알 수 있다. 이와는 반대로 일반모형(G)과 같은 비교적 복잡한 모형에서 자료를 생성하여, 이 모형보다는 단순한 대각모형(D), 랜덤절편항 모형(RI) 적합시켜 얻은 RMSE값을 살펴보면, $RMSE_{D,G} = 108.8$, $RMSE_{RI,G} = 131.9$ 로써 추정의 효율이 떨어짐을 확인할 수 있다.

RMSE 값 간의 차이를 살펴보면, $RMSE_{G,m_1}$ 와 $RMSE_{RI,m_1}$ 간의 차이는 크게 나타난다. 특히 $RMSE_{G,G} = 100.0$와 $RMSE_{RI,G} = 131.9$ 간의 차이와 $RMSE_{G,D} = 101.8$와 $RMSE_{RI,D} = 109.1$ 간의 차이가 크게 나타난다. 또한 $RMSE_{D,m_1}$과 $RMSE_{RI,m_1}$간의 차이도 크게 나타나며, 세부적으로 살펴보면 $RMSE_{D,G} = 108.8$와 $RMSE_{RI,G} = 131.9$ 간의 차이와 $RMSE_{D,D} = 100.0$와 $RMSE_{RI,D} = 109.1$ 간의 차이가 크다. 반면 $RMSE_{G,m_1}$와 $RMSE_{D,m_1}$간의 차이는 전체적으로는 작게 나타난다.

이상의 결과를 요약하면 주어진 자료에 대해서 소지역 추정시 랜덤절편항 모형(RI)을 가정하는 것은 될 수 있다면 피하는 것이 바람직하며, 그러나 반드시 복잡한 일반모형(G)만을 고집할 필요는 없겠고, 일반모형(G)이나 대각모형(D) 중 어떤 것을 선택하여도 무난할 것으로 판단된다. 더불어 소지역 추정의 효율을 높일 수 있는 추가적인 랜덤계수의 도입도 고려해 볼만 연구 사항이다.

회귀모형에 비해서 Multilevel 모형은 소지역들이 지역들 간의 특성을 보유하면서 하나의 통일된 형태로 모형이 표현된다는 것을 장점으로 들 수 있다. 각각의 소지역에서 표본의 수가 적을 때 Multilevel 모형을 이용하여 추정하였을 경우 좋은 결과를 예측하기가 어려울 것으로 생각할 수 있으나, 이러한 경우에도 Multilevel 모형을 이용한 추정이 소지역별로 분리되어 추정된 회귀모형 추정보다 평균적으로 더 좋은 결과를 보여 준다. 다음의 <표 2.2>

는 이러한 결과를 설명한다.

<표 2.2> RMSE 와 RARE

Estimator	Data Generation Model	
	General Model	Separate Regression Model
G	100.0 (100.0)	88.1 (83.1)
Separate Regression	247.6 (154.7)	100.0 (100.0)

Multilevel 모형의 자료 생성 시 앞에서 계산된 모수값을 이용하였고, 일반적 회귀모형에서 자료 생성시 모수의 추정값 $\hat{\sigma}^2$ 은 각각의 소지역에서 추정된 값을 이용하였다. 소지역 간의 랜덤효과를 고려하지 않은 일반적 회귀모형을 이용했을 경우 소지역 추정은 심각한 효율의 손실을 가져올 수 있음을 확인 할 수 있다. 한편, $RMSE_{SR,G}$ 값을 각각의 소지역별로 분리하여 살펴보면, 소지역의 표본크기가 클수록 작은 경향을 보인다. 소지역의 표본크기가 큰 경우에는 $RMSE_{G,G}$ 값과 $RMSE_{SR,G}$ 값 간의 차이는 줄어드는 경향을 보인다.

다음은 모의실험의 결과로부터 $\hat{\mu}_i$ 의 근사 MSE 성질을 설명하고자 한다. 비교기준으로 $\hat{\mu}_i$ 의 MSE 근사식은 다음 식을 이용하고,

$$MSE(\hat{\mu}_i) = E(\hat{\mu}_i - \mu_i)^2$$

$$= E(\tilde{\mu}_i - \mu_i)^2 + E(\hat{\mu}_i - \tilde{\mu}_i)^2 \ , i=1,2,\cdots,m$$

$$\approx \{ \overline{X}_i^t (G_i^{-1})^t \Omega \overline{X}_i \}$$

$$+ \left\{ \sigma^2 \overline{X}_i (G_i^{-1})^t Z_i \left(\sum_{i=1}^{m} Z_i^t G_i^{-1} X_i^t X_i Z_i \right)^{-1} Z_i^t G_i^{-1} \overline{X}_i \right\}$$

$$+\left\{ \overline{X_i}^t (G_i^{-1}) \left(\sum_{j=1}^{s-1} \sum_{k=1}^{s-1} b_{jk} \triangle_j C_i \triangle_k^t \right) G_i^{-1} \overline{X_i} \right.$$

$$\left. - 2\,\overline{X_i}^t (G_i^{-1})^t \left(\sum_{j=1}^{s-1} b_{j,s} \triangle_j \right) R_i \Omega \overline{X_i} + b_{ss}\, \overline{X_i}^t \Omega S_i \Omega \overline{X_i} \right\},$$

$$= T_1 + T_2 + T_3$$

모의실험에 의한 $\hat{\mu}_i$ 의 MSE 근사값으로는 다음의 식

$$MSE(\hat{\mu}_{iG}^{(r)}) = E(\hat{\mu}_{iG,G}^{(r)} - \mu_{iG}^{(r)})^2$$

을 이용한다. 단, $r = 1, 2, \cdots, R\,(=5000)$ 이고, T_1 , T_2 , T_3 의 X_i 값은 10%의 Simulation Subset 의 값으로 대체된다.

일반모형(G)에서 자료를 생성하여 일반모형(G)을 추정한 경우 MSE 의 근사는 매우 좋게 나타났다. 5000번의 모의실험 결과로부터 생성된 $MSE(\hat{\mu}_{iG}^{(r)})$의 값을 $MSE(\hat{\mu}_i)$ 값과 비교해 보면, 평균적인 과소추정의 양은 $MSE(\hat{\mu}_i)$값의 0.31%정도이고, 이러한 과소추정의 양 중 가장 큰 값은 $MSE(\hat{\mu}_i)$값의 약5.4%정도이며, 과대추정의 양 중 가장 큰 값은 $MSE(\hat{\mu}_i)$값의 약 4.8%정도로 나타났다.

모의실험 결과 T_1은 평균적으로 $MSE(\hat{\mu}_i)$의 94.6% 정도, T_3는 평균적으로 $MSE(\hat{\mu}_i)$의 4.3%정도를 차지한다. 각각의 소지역에서 살펴보면 T_1은 $MSE(\hat{\mu}_i)$의 87.4%에서 99.1%의 범위에 있고, T_3는 0.7%에서 10.5%의 범위에 있으며, T_2는 $MSE(\hat{\mu}_i)$값의 2.2%미만의 범위에 있음이 확인되었다.

마지막으로 $MSE(\hat{\mu}_i)$의 추정결과를 살펴보기로 한다. 5000번의 모의실험 결과로부터 계산된 $\widehat{MSE}(\hat{\mu}_{iG}^{(r)}) = \hat{T}_1^* + \hat{T}_2^* + 2\,\hat{T}_3^*$의 값을 $\widehat{MSE}(\hat{\mu}_i) = \hat{T}_1 + \hat{T}_2 + 2\,\hat{T}_3$의 값과 비교해 보면 근사적으로 Unbiased

되어 있고, $\widehat{MSE}(\hat{\mu}_{iG}^{(r)}) = \hat{T}_1^* + \hat{T}_2^*$값을 $\widehat{MSE}(\hat{\mu}_i) = \hat{T}_1 + \hat{T}_2$과 비교해 보면, 평균적으로 약 9.1%정도 과소추정됨이 확인된다. 이는 Singh et al.(1988), Prasad and Rao(1990)의 결과와도 일치한다.

소지역 추정시 특정 분산성분을 갖는 모형이 소지역 추정의 정확도를 개선할 수 있다는 사실을 제안한 연구는 Prasad and Rao(1990), Battese et. al(1981, 1988)에 의해서였다. Moura et al.(1999)는 특정 분산항을 갖는 모형인 Multilevel 모형이 소지역 추정의 정확도를 개선할 수 있다는 적용사례를 제시하여 소지역 추정에 관한 연구를 가일층 진보시켰다.

Moura et al.(1999)의 모의실험 결과로부터 확인할 수 있는 사항은 다음과 같이 요약된다. 첫째, 랜덤절편항 모형(RI)보다는 좀더 복잡한 구조를 갖는 대각모형(D), 일반모형(G)과 같은 분산성분 모형을 이용할 경우 소지역 추정의 정확도가 개선될 수 있다. 둘째, 분산성분을 갖는 모형을 가정할 때, 좀 더 복잡한 모형을 가정하더라도 추정의 효율은 심하게 떨어지지 않는다. 셋째, 소지역 공변량을 도입하여 소지역 추정의 정확도를 개선할 수 있다. 넷째, 소지역 추정시 Multilevel 모형 이용이 일상적인 회귀모형 이용보다 오히려 선호 되어야 한다.

한편, Moura et al(1999)의 모의실험 결과를 살펴보면, $\hat{\mu}_i$ 의 *MSE* 근사는 정도가 높게 나타나고 *MSE* 추정은 근사적으로 Unbiased 되어 있으나 추가적으로 근사의 정확한 order 에 대한 이론적인 연구가 계속 이루어져야 하고, 소지역 추정의 정확도를 높일 수 있는 일반혼합효과 모형의 적용 연구도 진행되어야 할 사항이다.

2.4.2 단위 수준(Unit-Level) 모형 적용사례

(1) 서 론

미 농림부에서는 농작물 재배면적의 예측을 보다 정확히 하기 위하여 지역

관측위성(LANDSAT)을 이용한 위성자료를 이용해오고 있다. 여기에서는 1978년 6월 Iowa 12개 County에 대하여 옥수수와 콩의 재배면적에 관한 사전 조사자료와, 8월과 9월 사이에 지역위성으로부터 관측된 옥수수와 콩의 실제 재배면적자료를 근거로하여 각 County에 대하여 옥수수와 콩의 재배면적에 관한 예측 문제를 가정된 단위수준(Unit-level) 모형에 근거하여 다뤘다. 이용된 자료는 미 농림부(USDA)에서 제시한 자료로써 <표 2.3>에 소개하였으며, 12개 County 중 37개 구역(Segment)의 조사자료만을 분석에 이용하였다.

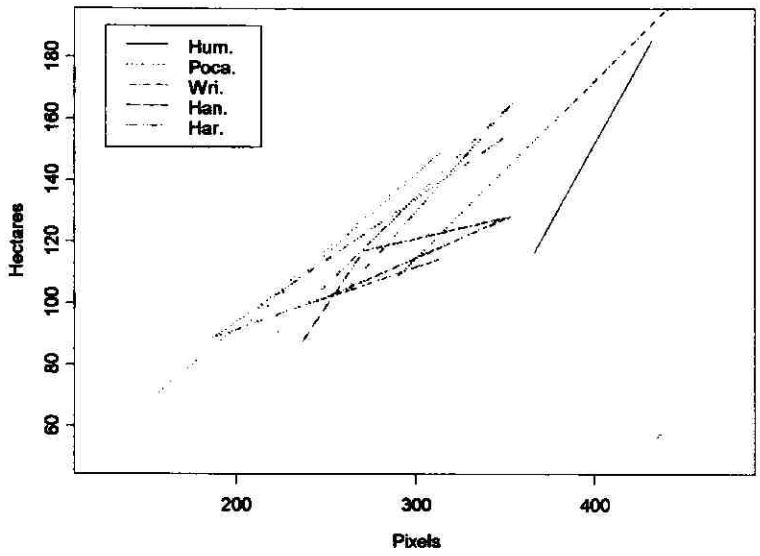
<표 2.3> Survey and Satellite Data for Corn and Soybeans in 12 Iowa Couties

카운티	No. of segment		Reported hectares		No. of pixels in sample segments		Mean No. of pixels per segment	
	Sample	County	Corn	Soybeans	Corn	Soybeans	Corn	Soybeans
Cerro Gordo	1	545	165.76	8.09	374	55	295.29	189.70
Hamilton	1	566	96.32	106.03	209	218	300.40	196.65
Worth	1	394	76.08	103.60	253	250	289.60	205.28
Humbolt	2	424	185.35	6.47	432	96	290.74	220.22
			116.43	63.82	367	178		
Franklin	3	564	162.08	43.50	361	137	318.21	188.06
			152.04	71.43	288	206		
			161.75	42.49	369	165		
Pocahontas	3	570	92.88	105.26	206	218	257.17	247.13
			149.94	76.49	316	221		
			64.75	174.34	145	338		
Winnebago	3	402	127.07	95.67	355	128	291.77	185.37
			133.55	76.57	295	147		
			77.70	93.48	223	204		
Wright	3	567	206.39	37.84	459	77	301.26	221.36
			108.33	131.12	290	217		
			118.17	124.44	307	258		
Webster	4	687	99.96	144.15	252	303	262.17	247.09
			140.43	103.60	293	221		
			98.95	88.59	206	222		
			131.04	115.58	302	274		
Hancock	5	569	114.12	99.15	313	190	314.28	198.66
			100.60	124.56	246	270		
			127.88	110.88	353	172		
			116.90	109.14	271	228		
			87.41	143.66	237	297		
Kossuth	5	965	93.48	91.05	221	167	298.65	204.61
			121.00	132.33	369	191		
			109.91	143.14	343	249		
			122.66	104.13	342	182		
			104.21	118.57	294	179		
Hardin	6	556	88.59	102.59	220	262	325.99	177.05
			88.59	***29.46***	***340***	***87***		
			165.35	69.28	355	160		
			104.00	99.15	261	221		
			88.63	143.66	187	345		
			153.70	94.49	350	190		

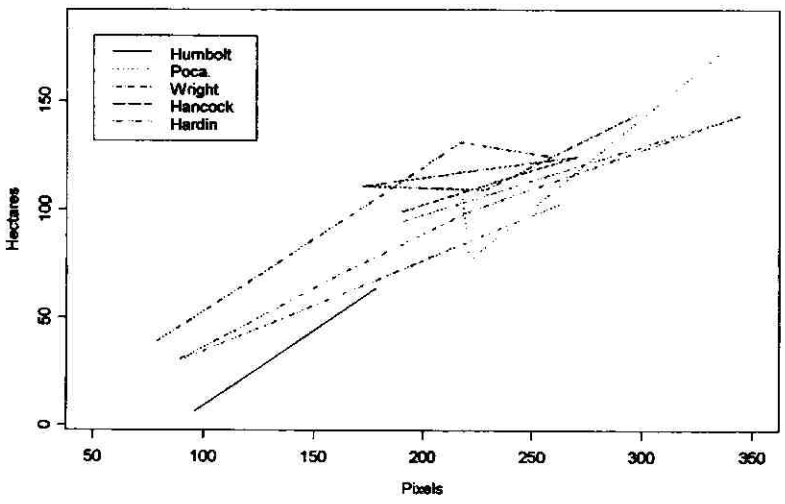
* Hardin의 자료중 하나의 관측점은 이상치로 간주하고 분석에서 제외됨

County별로 옥수수와 콩의 재배면적에 관한 조사자료와 위성에 의해 관측된 Pixel수에 관한 상관성을 살펴보기 위해 각각에 대해 산점도를 그려 직선으로 연결해 보면 <Fig1>과 <Fig2>와 같다.

<Fig 1> Corn ha vs pixels by county



<Fig 2> Soybean ha vs pixels by county



성이 있는 것으로 가정하는 것이 합당하며, 또한 Model의 랜덤오차들은 Nested-Error Model에 의해 정의되는 것으로 가정한다.

(2) 분산 성분 모형(Components-of-variance Models)

가정된 분산성분 모형은 다음의 (2.49)과 같다.

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_{ij} \tag{2.49}$$

단, $i = 1, 2, \cdots, T$ (county 수, $T = 12$), $j = 1, 2, \cdots, n_i$ (i번째 county에서 sample segment 수), y_{ij}= i번째 county에서 j번째 sample segment에 있는 옥수수(또는 콩)의 조사된 재배면적(ha), x_{1ij}= i번째 county에서 j번째 segment에 있는 옥수수의 pixel 수, x_{2ij}= 콩의 pixel 수, u_{ij}는 다음과 같이 가정한다.

$$u_{ij} = \nu_i + \varepsilon_{ij} \tag{2.50}$$

여기에서 ν_i= i번째 county 효과 $\overset{iid}{\sim} N(0, \sigma_\nu^2)$ ($i = 1, 2, \cdots, T$),

ε_{ij}= i번째 county에서 j번째 sample segment와 관련된 랜덤 효과 $\overset{iid}{\sim}$ $N(0, \sigma_e^2)$를 가정하며, 이때 공분산 구조는 다음의 (2.51)과 같다.

$$E(u_{ij} u_{pq}) = \begin{cases} \sigma_\nu^2 + \sigma_e^2 & , i = p, j = q \\ \sigma_\nu^2 & , i = p, j \neq q \\ 0 & , i \neq p \end{cases} \tag{2.51}$$

분산성분 모형 (2.49)과 (2.50)는 county내의 sample segment에서 옥수수와 콩의 상관구조를 완전하게 정의하지는 못한다. 그러나 이 연구에서는 사용된 자료가 다변량 모형을 적용하였을 경우 추정의 정도를 개선하지 못했기 때문에 단순히 일변량의 경우로 제한하였다.

i번째 county에서 옥수수(또는 콩) 재배면적의 표본 평균은

$\bar{y}_{i\cdot} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$로 주어지고, (2.49)식과 (2.50)식을 이용하여 표현하면 다음과 같다.

$$\bar{y}_{i\cdot} = \beta_0 + \beta_1 \bar{x}_{1i\cdot} + \beta_2 \bar{x}_{2i\cdot} + \nu_i + \bar{e}_{i\cdot} \qquad (2.52)$$

여기에서 $\bar{x}_{1i\cdot} = \frac{\sum_{j=1}^{n_i} x_{1ij}}{n_i}$, $\bar{x}_{2i\cdot} = \frac{\sum_{j=1}^{n_i} x_{2ij}}{n_i}$, $\bar{e}_{i\cdot} = \frac{\sum_{j=1}^{n_i} e_{ij}}{n_i}$로 주어진다.

모평균을 y_i 라 표현한다면,

$$y_i = \beta_0 + \beta_1 \bar{x}_{1i(p)} + \beta_2 \bar{x}_{2i(p)} + \nu_i \ , \qquad (2.53)$$

여기에서 $\bar{x}_{1i(p)} = \frac{\sum_{j=1}^{N_i} x_{1ij}}{N_i}$, $\bar{x}_{2i(p)} = \frac{\sum_{j=1}^{N_i} x_{2ij}}{N_i}$ 로 주어지고, $N_i =$ i번째 county에서 segment수의 총합을 나타낸다. 위성자료로부터 $\bar{x}_{1i(p)}$, $\bar{x}_{2i(p)}$ 가 계산된다.

(2.53)식을 기반으로 평균 농작물 재배면적에 관한 추정 문제를 다루고자 한다. 유한 모집단 모형에서 i번째 county에서 옥수수(또는 콩)의 평균 재배면적은 $\bar{Y}_{i\cdot} = \frac{\sum_{j=1}^{N_i} Y_{ij}}{N_i}$이고, (2.53)식의 y_i 는 표본 추출률이 작을 때 $\bar{Y}_{i\cdot}$에 대한 예측변수를 나타낸다.

(2.49) ~ (2.51)식을 행렬형식으로 표현하면, $Y_i = (y_{i1}, y_{i2}, \cdots, y_{in_i})^t$, $Y = (Y_1^t, Y_2^t, \cdots, Y_T^t)$ 로부터

$$Y = X\beta + u \ , \qquad (2.54)$$

여기에서 y_{ij} 에 대응되는 X의 행은 $x_{ij} = (1, x_{1ij}, x_{2ij})$,

$\beta = (\beta_0, \beta_1, \beta_2)^t$, u 의 공분산 행렬은

$$E(uu^t) = V = block\ diag(V_1, V_2, \cdots, V_T), \qquad (2.55)$$

$$V_i = J_i\,\sigma_\nu^2 + I_i\,\sigma_e^2\ , \qquad (2.56)$$

여기에서 J_i는 order가 n_i 이며 모든 원소가 1인 정방행렬이고, I_i는 order가 n_i인 단위행렬을 나타낸다.

위의 (2.53)식을 행렬로 표현하면 다음과 같다.

$$y_i = \bar{x}_{i(p)}\beta + \nu_i\ , \qquad (2.57)$$

여기에서 $\bar{x}_{i(p)} = \dfrac{\sum_{j=1}^{N_i} x_{ij}}{N_i} = (1,\ \bar{x}_{1i(p)},\ \bar{x}_{2i(p)})$로 주어진다.

(3) 모수 추정

앞에서 언급한 (2.49), (2.50), (2.51)을 충족한다는 가정하에서 모수에 대한 추정을 고려해 보자. 만약에 u_{ij} 가 기지이면, ν_i 의 best predictor는 $E(\nu_i \mid \bar{u}_{i\cdot})$가 되고, 여기에서 $\bar{u}_{i\cdot} = \dfrac{\sum_{j=1}^{n_i} u_{ij}}{n_i}$이다.

(2.49)과 (2.50)의 가정하에서 ν_i 와 $\bar{u}_{i\cdot}$는 다음과 같은 이변량 정규분포를 따르게 되고,

$$(\nu_i,\ \bar{u}_{i\cdot}) \sim N\left(0, \begin{pmatrix} \sigma_\nu^2 & \sigma_\nu^2 \\ \sigma_\nu^2 & \sigma_\nu^2 + \dfrac{\sigma_e^2}{n_i} \end{pmatrix}\right),$$

주변확률 분포로부터 $E(\nu_i \mid \bar{u}_{i\cdot}) = \bar{u}_{i\cdot}\dfrac{\sigma_\nu^2}{\sigma_\nu^2 + \dfrac{\sigma_e^2}{n_i}} = \bar{u}_{i\cdot}\,g_i$ 가 주어

진다. 여기에서 $g_i = m_i^{-1}\ \sigma_\nu^2$ 이고 $m_i = (\ \sigma_\nu^2 + \frac{\sigma_e^2}{n_i}\ )$ 이다.

따라서 위의 관계로부터 다음의 결과를 얻을 수 있다.

$$E(\nu_i - \bar{u}_{i\cdot}\, g_i \mid \bar{u}_{i\cdot}) = 0\ ,$$

$$E\{(\nu_i - \bar{u}_{i\cdot}\, g_i)^2\} = \sigma_\nu^2 (1 - g_i)$$

$$= n_i^{-1}\ \sigma_e^2 - n_i^{-2}\ \sigma_e^2\ m_i^{-1}\ \sigma_e^2 \qquad (2.58)$$

여기에서 σ_e^2 과 σ_ν^2은 미지인 값이며, 만약에 σ_e^2, σ_ν^2이 기지이면 β의 GLS 추정량은 다음과 같이 주어진다.

$$\hat{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} Y \qquad (2.59)$$

이때 i번째 county의 효과 ν_i 는 다음 식으로 추정된다.

$$\tilde{\nu}_i = \tilde{u}_{i\cdot}\, g_i\ , \qquad (2.60)$$

단, $\tilde{u}_{i\cdot} = \dfrac{\sum_{j=1}^{n_i} \tilde{u}_{ij}}{n_i}$, $\tilde{u}_{ij} = y_{ij} - x_{ij}\hat{\beta}$ 로 주어진다. 여기에 대응되는 y_i (i번째 county의 평균 재배면적) BLUP 추정량은 다음의 (2.61)식으로 주어지고,

$$\tilde{y}_i = \bar{x}_{i(\beta)}\hat{\beta} + \tilde{\nu}_i \qquad (2.61)$$

(2.61)식으로부터 다음의 결과를 얻을 수 있다.

$$E\{(\tilde{y}_i - y_i)^2\} = \sigma_\nu^2 (1 - g_i) + c_i V(\hat{\beta})\ c_i^t \qquad (2.62)$$

단, $V(\hat{\beta}) = (X^t V^{-1} X)^{-1}$,

$$c_i = \bar{x}_{i(\beta)} - g_i \bar{x}_{i\cdot}\,,\ \ \bar{x}_{i\cdot} = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} = (1,\ \bar{x}_{1i\cdot}\,,\ \bar{x}_{2i\cdot})$$

$$x_{ij} = (1, x_{1ij}, x_{2ij}),$$

$$\bar{x}_{i(p)} = \frac{\sum_{j=1}^{N_i} x_{ij}}{N_i} = (1,\ \bar{x}_{1i(p)},\ \bar{x}_{2i(p)})$$

(2.62)식은 (2.58)식 보다 $c_i V(\hat{\beta})\ c_i^t$ 만큼 큰 양을 나타낸다.

유한 모집단에서 i번째 county에서 재배면적의 모평균에 대한 추정은 다음 식을 이용한다.

$$\frac{\sum_{j=1}^{n_i} y_{ij} + \sum_{j=n_i+1}^{N_i} (x_{ij}\hat{\beta} + \tilde{\nu}_i)}{N_i} \qquad (2.63)$$

이 추정식은 표본 추출률이 작을 경우에는 (2.61)식과 근사적으로 같다. 이러한 사실을 근거로 해서 (2.61)식을 이용하여 문제에 접근하였다.

대부분의 문제에서는 σ_ν^2과 σ_e^2은 미지이므로 추정되어야 한다. β 의 GLS 추정량 $\hat{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} Y$ 과 $\tilde{y}_i = \bar{x}_{i(p)}\hat{\beta} + \tilde{\nu}_i$ 는 σ_ν^2과 σ_e^2이 미지이므로 추정되어야 한다.

식(2.49)에서 가정한 모형으로부터

$$\hat{\sigma}_e^2 = \hat{e}^t \hat{e} \Big(\sum_{i=1}^{T} (n_i - 1) - 2 \Big), \qquad (2.64)$$

여기에서 $\hat{e}^t \hat{e}$ 는 $(y_{ij} - \bar{y}_{i.})^2$ 의 합이며, $n_i > 1$일때 (2.49)과 (2.50)의 가정하에서 $\hat{\sigma}_e^2$은 σ_e^2에 대해서 불편이고, $d_e \frac{\hat{\sigma}_e^2}{\sigma_e^2} \sim \chi^2(d_e)$ 인 관계가 성립하므로 $d_e = \sum_{i=1}^{T}(n_i - 1) - 2$ 로 주어진다.

county 효과 $\hat{\sigma}_\nu^2$ 은 i county 에 대해서 다음의 (2.65)식의 잔차를 고려하여 구해진다.

$$\breve{u}_{i.} = \bar{y}_{i.} - \bar{x}_{i.} (X^t X)^{-1} X^t Y \ , \qquad (2.65)$$

여기에서

$$E(\ \breve{u}_{i\cdot}^{\ 2}) = b_i\,\sigma_\nu^2 + d_i\,\sigma_e^2\ , \tag{2.66}$$

단, $b_i = 1 - 2n_i\,\bar{x}_{i\cdot}\,(X^tX)^{-1}X^t\ \bar{x}_{i\cdot}^{\ t}$

$$+\ \bar{x}_{i\cdot}\,(X^tX)^{-1}\Big(\sum_{j=1}^{T} n_j^2\ \bar{x}_{j\cdot}^{\ t}\,\bar{x}_{j\cdot}\Big)(X^tX)^{-1}\ \bar{x}_{i\cdot}^{\ t}$$

$$d_i = n_i^{-1}\left\{1 - n_i\,\bar{x}_{i\cdot}\,(X^tX)^{-1}\ \bar{x}_{i\cdot}^{\ t}\right\}$$

county 들에 대한 잔차의 가중제곱합의 평균은 다음의 (2.67)식으로 주어지고,

$$\hat{m}\,\cdot\,\cdot = \frac{\sum_{i=1}^{T} n_i\ \breve{u}_i^{\,2}}{\sum_{i=1}^{T} n_i b_i}\ , \tag{2.67}$$

위의 (2.67)식의 기대값은 다음식과 같이 표현된다.

$$E(\ \hat{m}\,\cdot\,\cdot) = E\left(\frac{\sum_{i=1}^{T} n_i\ \breve{u}_i^{\,2}}{\sum_{i=1}^{T} n_i b_i}\right)$$

$$= m\,\cdot\,\cdot$$

$$= \sigma_\nu^2 + c\,\sigma_e^2$$

$$\text{단,}\quad c = \frac{\sum_{i=1}^{T} n_i d_i}{\sum_{i=1}^{T} n_i b_i}$$

(2.49)과 (2.50)의 가정하에서 $\hat{m}\,\cdot\,\cdot$ 와 $\hat{\sigma}_e^2$은 서로 독립이므로 σ_ν^2의 추정량은 다음의 결과를 이용한다.

$$\hat{\sigma}_\nu^2 = \max\left\{\ \hat{m}\,\cdot\,\cdot - c\ \hat{\sigma}_e^2\ ,\ 0\right\} \tag{2.68}$$

이상의 추정에 근거하면 $g_i = \dfrac{\sigma_\nu^2}{\sigma_\nu^2 + \dfrac{\sigma_e^2}{n_i}}$ 의 추정식 $\widetilde{g_i}$ 는 다음의

(2.69)식과 같고,

$$\widetilde{g_i} = \frac{\widehat{\sigma}_\nu^2}{\widehat{\sigma}_\nu^2 + \dfrac{\widehat{\sigma}_e^2}{n_i}} \tag{2.69}$$

따라서 (2.57)식 $y_i = \bar{x}_{i(p)}\beta + \nu_i$, $\bar{x}_{i(p)} = \dfrac{\sum_{j=1}^{N_i} x_{ij}}{N_i} = (1, \bar{x}_{1i(p)}, \bar{x}_{2i(p)})$

는 다음 식으로 추정될 수 있다.

$$\widehat{y_i} = \bar{x}_{i(p)}\widehat{\beta} + \hat{u}_{i\cdot}\,\hat{g}_{i\cdot}\,, \tag{2.70}$$

여기에서 $\widehat{\beta}$ 은 β 의 GLS 추정량인 $\widehat{\beta}$ 에서 V를 $\widehat{V}$ 로 대체하고 다음과 같이 구할 수 있다.

$$\widehat{\beta} = (X^t\,\widehat{V}^{-1}X)^{-1}X^t\,\widehat{V}^{-1}Y\,,$$

단, $\widehat{V}(=\widehat{E(uu^t)}) = block\ diag(\widehat{V}_1, \widehat{V}_2, \cdots, \widehat{V}_T)$, $\widehat{V}_i = J_i\,\widehat{\sigma}_\nu^2 + I_i\,\widehat{\sigma}_e^2$,

$\hat{u}_{i\cdot} = \bar{y}_{i\cdot} - \bar{x}_{i\cdot}\,\widehat{\beta}$, $\widehat{g_i} = g_i$ 에 대해서 불편인 추정량이다.

추정오차 $y_i - \widehat{y_i}$ 의 분산에 대한 추정식은 다음의 결과를 이용한다(다변량의 경우에는 Fuller and Harter(1987)를 참조).

$$\begin{aligned}\widehat{y_i} &= \bar{x}_{i(p)}\widehat{\beta} + \hat{u}_{i\cdot}\,\widehat{g_i} \\ &= \bar{x}_{i(p)}\widehat{\beta} + (\bar{y}_{i\cdot} - \bar{x}_{i\cdot}\,\widehat{\beta})\,\widehat{g_i}\,,\end{aligned} \tag{2.71}$$

여기에서 $\widehat{g_i} = 1 - \widehat{h_i}$,

$$\widehat{h_i} = \{\widehat{m_i} + \widehat{k_i} + (n_i^{-1} - c)^2\,\widehat{w_i}\}^{-1}\{n_i^{-1}\,\widehat{\sigma}_e^2 + (n_i^{-1} - c)\,n_i^{-1}\,\widehat{w_i}\},$$

단, $\widehat{m}_i = \widehat{m}\,.\,. + (n_i^{-1} - c)\,\widehat{\sigma}_e^2$,

$\widehat{w}_i = 2\,d_e^{-1}\,\widehat{m}_i^{-1}\,\widehat{\sigma}_e^4$,

$\widehat{k}_i = 2\,\widehat{\sigma}_e^2(\dot{\sigma}_{ff} + n_i^{-1})^{-1}(\sum_{j=1}^{T} n_j b_j)^{-2}(\sum_{j=1}^{T} n_j^2 b_j(\dot{\sigma}_{ff} + n_j^{-1})^2)$,

단, $\dot{\sigma}_{ff} = \max\{0\,,\ (T-5)^{-1}(T-3)\,\widehat{\sigma}_e^{-2}\,\widehat{m}\,.\,. - c\}$,

$b_j = 1 - 2n_j\,\bar{x}_{j\cdot}\,.\,(X^tX)^{-1}X^t\,\bar{x}_{j\cdot}^{\,t}$

$+\ \bar{x}_{j\cdot}\,.\,(X^tX)^{-1}(\sum_{i=1}^{T} n_i^2\,\bar{x}_{i\cdot}^{\,t}\,\bar{x}_{i\cdot}\,.)(X^tX)\,\bar{x}_{j\cdot}^{\,t}$,

$$c = \frac{\sum_{i=1}^{T} n_i d_i}{\sum_{i=1}^{T} n_i b_i}$$,

$d_i = n_i^{-1}(1 - n_i\,\bar{x}_{i\cdot}\,.\,(X^tX)^{-1}\,\bar{x}_{i\cdot}^{\,t})$,

$d_e = 22$

(2.71)식의 결과를 이용하여 추정오차의 분산의 추정식을 계산하면 다음과 같이 주어진다.

$\widehat{Var}(\hat{y}_i - y_i) = n_i^{-1}\,\widehat{\sigma}_e^2 - \hat{\phi}_i + \hat{c}_i\hat{V}(\hat{\beta})\,\hat{c}_i^t + \hat{h}_i^2\hat{k}_i + d_e^{-1}\,\hat{r}_i^2\hat{\phi}_i + d_e^{-1}\,\hat{r}_i^2\hat{h}_i\,\widehat{\sigma}_e^2$

단, $\hat{c}_i = \bar{x}_{i(\beta)} - \hat{g}_i\,\bar{x}_{i\cdot}$,

$\hat{\phi}_i = (d_e+1)^{-1}d_e\ddot{\phi}_i - d_e^{-1}\,n_i^{-1}\,\widehat{\sigma}_e^2\hat{h}_i$,

$\ddot{\phi}_i = n_i^{-2}(\widehat{\sigma}_e^2 + (n_i^{-1} - c)\widehat{w}_i)^2(\widehat{m}_i + \hat{k}_i + (n_i^{-1} - c)^2\,\widehat{w}_i)^{-1}$,

$\hat{r}_i = 1 - (1 - n_i c)\,\hat{h}_i$

(4) 추정 결과

i) $\widehat{\sigma}_v^2 = \max\{\widehat{m}\,.\,. - c\,\widehat{\sigma}_e^2\ ,\ 0\}$

$$\widehat{\sigma_e^2} = \hat{e}^t \hat{e} \left(\sum_{i=1}^{T} (n_i - 1) - 2 \right), \quad n_i > 1$$

ii) β 의 GLSE

$$\hat{\beta} = (X^t \hat{V}^{-1} X)^{-1} X^t \hat{V}^{-1} Y,$$

단, $\hat{V} = block\ diag(\hat{V}_1, \hat{V}_2, \cdots, \hat{V}_T)$

$$\hat{V}_i = J_i \widehat{\sigma_\nu^2} + I_i \widehat{\sigma_e^2}$$

iii) $\hat{y}_i = \bar{x}_{i(\beta)} \hat{\beta} + \hat{u}_{i.} \hat{g}_i$

$= \bar{x}_{i(\beta)} \hat{\beta} + (\bar{y}_{i.} - \bar{x}_{i.} \hat{\beta}) \hat{g}_i$

옥수수의 경우

$$\hat{y}_{ij} = 51 + 0.329 x_{1ij} - 0.134 x_{2ij},$$

(25) (0.050) (0.056)

$$\widehat{\sigma_e^2} = 150, \quad \widehat{\sigma_\nu^2} = 140$$

(45) (89)

콩의 경우

$$\hat{y}_{ij} = -16 + 0.028 x_{1ij} + 0.494 x_{2ij}$$

(29) (0.058) (0.065)

$$\widehat{\sigma_e^2} = 195, \quad \widehat{\sigma_\nu^2} = 272$$

(59) (49)

$$c = 0.349$$

<표 2.4> 옥수수의 예측 재배면적과 표준오차

County	Sample segment	Predicted hectares	standard error		
			Best predictor	Survey regression predictor	Sample mean
Cerro Gordo	1	122.2	9.6	13.7	30.5
Hamilton	1	126.3	9.5	12.9	30.5
Worth	1	106.2	9.3	12.4	30.5
Humboldt	2	108.0	8.1	9.7	21.5
Franklin	3	145.0	6.5	7.1	17.6
Pocahontas	3	112.6	6.6	7.2	17.6
Winnebago	3	112.4	6.6	7.2	17.6
Wright	3	122.1	6.7	7.3	17.6
Webster	4	115.8	5.8	6.1	15.2
Hancock	5	124.3	5.3	5.7	13.6
Kossuth	5	106.3	5.2	5.5	13.6
Hardin	5	143.6	5.7	6.1	13.6

* Best Predictor : $\hat{y}_i = \bar{x}_{i(p)}\hat{\beta} + \hat{u}_{i.}\hat{g}_i$

* Survey Regresson Predictor : $y_i = \bar{y}_{i.} + (\bar{x}_{i(p)} - \bar{x}_{i.})\hat{\beta}$

* 표본평균의 추정 표준오차는 county내의 mean square를 county내의 segment의 수로 나누어 root를 취한 값임.

<표 2.5> 콩의 예측 재배면적과 표준오차

County	Sample segment	Predicted hectares	standard error		
			Best predictor	Survey regression predictor	Sample mean
Cerro Gordo	1	77.8	12.0	15.6	29.1
Hamilton	1	94.8	11.8	14.8	29.1
Worth	1	86.9	11.5	14.2	29.1
Humboldt	2	79.7	9.7	11.1	20.6
Franklin	3	65.2	7.6	8.1	16.8
Pocahontas	3	113.8	7.7	8.2	16.8
Winnebago	3	98.5	7.7	8.3	16.8
Wright	3	112.8	7.8	8.4	16.8
Webster	4	109.6	6.7	7.0	14.6
Hancock	5	101.0	6.2	6.5	13.0
Kossuth	5	119.9	6.1	6.3	13.0
Hardin	5	74.9	6.6	6.9	13.0

이상의 추정결과를 요약하면 다음과 같이 정리할 수 있다. 첫째, sample segment가 증가할수록 표본평균의 표준오차는 감소하는 경향을 볼 수 있다. 둘째, 표본평균의 표준오차는 survey regression predictor의 표준오차보다 상당히 크다는 사실을 확인할 수 있다.

$$\left(\frac{\textit{best predictor}\text{의 표준오차}}{\textit{survey regression predictor}\text{의 표준오차}} = 0.77 \sim 0.97\right).$$

또한 sample segment의 수가 3 이하일 때 Best Predictor는 작은 표준오차를 가지며, Predictor의 정확도는 sample segment의 수가 한 county에서 3에서 4 또는 5정도일때 정확도가 좋게 나타났다.

Survey Regression Predictor는 12개의 county의 평균 농작재배면적에 대해

서 unbiased되어있고, 상대적으로 작은 분산을 갖고 있다. 따라서 Survey Regression Predictor는 전체 지역에 대해서 적절한 Predictor라고 볼수 있다. 각각의 county에 대한 예측은 적당한 가중합을 이용하여 전지역에 대하여 불편인 특성을 갖는 Survey Regression Predictor와 같게 되도록 수정하는것도 생각해 볼 문제이다.

결론적으로 보조 변수로써 위성관측 자료를 이용한 분산성분 모형 적용은 소지역에서 농작 재배면적을 예측하는데 유용한 결과를 제공한다는 사실을 확인할 수 있다.

2.4.3 일반 통계 모형을 적용한 추정 사례

(1) 서 론

갤럽은 미국 전역의 알콜중독과 마약복용률을 추정하기 위해 주(State) 단위의 광범위한 가구조사를 실시하고 있다. 추정값은 주(State) 단위에서는 비교적 신뢰할만한 수준이나 sub-state 그룹에서의 추정은 표본이 불균형적으로 배정되거나 표본크기가 작을 경우에는 추정의 신뢰도가 상당히 떨어진다. 주(State) 단위로 조사된 조사 자료를 기반으로 sub-state에서의 추정을 살펴보며, 비교 추정량으로 직접 추정량, 합성추정량, 복합추정량, 경험적 베이즈 추정량이 제시된다.

$n_i(i=1, 2, \cdots, I)$를 i번째 계획구역에 할당된 표본크기로 하자 ($n=\sum_{i=1}^{I} n_i$). 각각의 계획구역에서의 표본은 RDD(random digit dialing)전화조사로 독립적으로 추출된다. 표본이 관측된 후 각각의 구역은 K개의 인구통계적 그룹으로 사후 층화된다. 이러한 그룹들은 성별(남, 여), 나이(18-24, 25-44, 45-64, 65세 이상)로 교차 분류되며 총 $K=2\times4=8$개의 그룹으로 분류된다.

J_i 를 i번째 계획구역에 있는 county 개수라 하고, n_{ijk}를 i번째 계획구역에 속해있는 j번째 county에서 k번째 인구통계적 그룹에 있는 관측값의 개수라 하자. 또한 S_{ij}는 i번째 계획구역 내의 j번째 county에 있는 인구통계적 그룹의 set을 나타내고, y_{ijkl}는 i번째 계획구역에 속해있는 j번째 county의 k번째 인구통계적 그룹에 대한 l번째 관측값(0 or1), w_{ijkl}는 표본조사로부터 얻은 표본추출가중치라 하자($i=1, 2, \cdots, I$; $j=1, 2, \cdots, J_i$; $k \in S_{ij}$; $l=1, 2, \cdots, n_{ijk}$). 이때 i번째 계획구역내에 있는 j번째 county에 대한 알콜중독 또는 마약복용률을 나타내는 π_{ij}를 추정하고자 한다.

(2) 직접추정량과 합성추정량

j번째 county에 대한 π_{ij} 의 직접추정값은 다음 (2.72)식으로 계산하고,

$$\widehat{\pi}_{ij}^{D} = \frac{\sum_{k \in S_{ij}} \sum_{l=1}^{n_{ijk}} w_{ijkl}\, y_{ijkl}}{\sum_{k \in S_{ij}} \sum_{l=1}^{n_{ijk}} w_{ijkl}} \qquad (2.72)$$

k번째 인구통계적 그룹에 대한 π_{ik}의 직접추정값은 다음 (2.73)식으로 계산된다.

$$\widehat{\pi}_{ik}^{D} = \frac{\sum_{j:\, k \in S_{ij}} \sum_{l=1}^{n_{ijk}} w_{ijkl}\, y_{ijkl}}{\sum_{j:\, k \in S_{ij}} \sum_{l=1}^{n_{ijk}} w_{ijkl}} \qquad (2.73)$$

여기에서 $j: k \in S_{ij}$ 는 인구통계적 그룹 k가 관측된 모든 county j에 대한 합을 의미한다.

이 추정량은 하나의 county로부터 이용할 수 있는 표본의 크기가 매우 작을 경우 추정값의 신뢰도는 매우 떨어진다. 또한, 하나의 county내에 있는 모든

관측점이 0이라면, 추정값도 0이 되고, 추정값의 표준오차도 0이 되어 추정의 신뢰도가 떨어진다. 따라서 직접추정량에서는 이러한 문제가를 개선되어야 할 필요가 있다.

합성추정량은 전화조사 자료와 U.S 센서스 국으로부터 획득한 보조자료의 관계를 이용하여 추정한다. county수준에서의 과거 알콜중독률을 이용한 합성추정량은 다음과 같다.

$$\widehat{\pi}_{ij}^{S1} = \sum_{k=1}^{K} a_{ijk} \quad \widehat{\pi}_{k}^{D} \tag{2.74}$$

여기에서 $\widehat{\pi}_{k}^{D}$ = k번째 인구통계 그룹에 대한 알콜중독률의 직접조사추정량, a_{ijk} = i번째 계획구역의 j번째 county의 k번째 인구통계적 그룹에 속해있는 개체들의 비율이다(최근의 센서스 추정치로부터 얻음). $\widehat{\pi}_{ij}^{S1}$에서 나타나 있듯이 k번째 인구통계적 그룹에 대한 알콜중독률과 마약복용률은 모든 county들에 대해서 동일한 것으로 가정되었다.

좀더 덜 제한적인 알콜중독과 마약복용률에 대한 합성추정량은 다음과 같다.

$$\widehat{\pi}_{ij}^{S2} = \sum_{k=1}^{K} a_{ijk} \quad \widehat{\pi}_{ik}^{D} \tag{2.75}$$

여기에서 $\widehat{\pi}_{ik}^{D}$는 i번째 계획구역에 있는 k번째 인구통계적 그룹에 대한 알콜중독 또는 마약복용률 π_{ik}의 직접조사추정량을 나타낸다. $\widehat{\pi}_{ij}^{S2}$에서는 k번째 그룹에 대한 알콜중독률과 마약복용률은 하나의 계획구역에 있는 모든 county에 대해서 동일한 것으로 가정되었고, 이러한 가정은 $\widehat{\pi}_{ij}^{S1}$에서보다는 합리적이고 덜 제한적인 가정이라 할 수 있다.

(3) 복합 추정량

직접조사추정량과 합성추정량의 절충이 복합추정량이다. 여기서 제안하는

복합추정량은 다음의 항등식에 근거한다.

$$\pi_{ij} = \sum_{k \in S_{ij}} a_{ijk} \pi_{ijk} + \sum_{k \notin S_{ij}} a_{ijk} \pi_{ijk} \qquad (2.76)$$

여기에서 π_{ijk} = 알콜중독 또는 마약복용률, a_{ijk} = i번째 계획구역의 j번째 county에 있는 k번째 인구통계적 그룹에 속해있는 개체들의 비율을 나타낸다.

π_{ij}의 단순복합추정량은 $k \in S_{ij}$에 대해서 π_{ijk}는 $\widehat{\pi}_{ijk}^{D}$로, $k \notin S_{ij}$에 대해서 π_{ijk}는 $\widehat{\pi}_{ik}^{D}$로 추정되며 다음과 같다.

$$\widehat{\pi}_{ij}^{C} = \sum_{k \in S_{ij}} a_{ijk} \widehat{\pi}_{ijk}^{D} + \sum_{k \notin S_{ij}} a_{ijk} \widehat{\pi}_{ik}^{D} \qquad (2.77)$$

(5.46)식에서 π_{ijk} ($k \in S_{ij}$)는 소표본에서 추정하므로 추정의 정확도가 떨어지게 되어, 따라서 최근의 보조정보를 이용하여 추정한다면 $\widehat{\pi}_{ijk}^{D}$ 에 대한 개선의 여지가 남아 있다. 이를 위하여 π_{ij}의 경험적 베이즈 추정량을 제안한다.

(4) 경험적 베이즈(EB) 추정량

경험적 베이즈 추정량을 산출하기에 앞서 다음을 가정하기로 한다.

i) π_{ijk} 들이 주어진 상태에서 y_{ijk}들은 서로 uncorrelate 되어 있고, 아래 조건을 만족한다.

$$E(y_{ijkl} \mid \pi_{ijk}) = \pi_{ijk} ,$$

$$Var(y_{ijkl} \mid \pi_{ijk}) = \pi_{ijk}(1 - \pi_{ijk}).$$

ii) π_{ijk} 들은 서로 uncorrelate 되어 있으며, 다음 식을 만족한다.

$$E(\pi_{ijk}) = \mu_{ik} ,$$

$$Var(\pi_{ijk}) = d\, \mu_{ik}^{2} .$$

만약 $\pi_{ijk} \sim U(0, 2\mu_{ik})$ 라면, 위에서 $d = \frac{1}{3}$. 즉, $\widehat{\pi_{ij}}^{S2}$에서의 가정 ($\pi_{ijk} = \mu_{ik}$)과는 달리, 특정한 인구통계적 그룹에 대해서 한 구역 내의 county들 간의 비율의 변동이 반영된다.

첫 번째 가정은 π_{ijk}가 주어진 상태에서 $\widehat{\pi_{ijk}}^{D}$는 서로 uncorrelate 되어 있음을 의미한다. 이때,

$$E(\widehat{\pi_{ijk}}^{D} \mid \pi_{ijk}) = \pi_{ijk} ,$$

$$Var(\widehat{\pi_{ijk}}^{D} \mid \pi_{ijk}) = c_{ijk}\, \pi_{ijk}(1-\pi_{ijk}) ,$$

$$\text{단, } c_{ijk} = \frac{\sum_{l=1}^{n_{ijk}} w_{ijkl}^{2}}{(\sum_{l=1}^{n_{ijk}} w_{ijkl})^{2}} .$$

손실함수로 제곱오차손실함수(squared error loss function)가 사용될 경우, π_{ij}의 선형 베이즈 추정량(linear Bayes estimator)은 아래 (2.78)식과 같이 주어진다.

$$\widehat{\pi_{ij}}^{B} = \sum_{k \in S_{ij}} a_{ijk}(B_{ijk}\ \widehat{\pi_{ijk}}^{D} + (1-B_{ijk})\mu_{ik}) + \sum_{k \notin S_{ij}} a_{ijk}\mu_{ik} , \quad (2.78)$$

$$\text{단, } B_{ijk} = \frac{d\ \mu_{ik}^{2}}{d\ \mu_{ik}^{2} + c_{ijk}(\mu_{ik} - (d+1)\ \mu_{ik}^{2})}$$

위의 베이즈 추정량은 미지인 모수 μ_{ik}를 포함하고 있기 때문에 μ_{ik}가 먼저 추정되어야만 한다. μ_{ik}가 $\widehat{\mu_{ik}}$ ($= \widehat{\pi_{ik}}^{D}$)로 대체되어 얻어진다면, π_{ij}의 경험적 베이즈 추정량 $\widehat{\pi_{ij}}^{EB}$는 다음과 같이 주어질 수 있다.

$$\widehat{\pi_{ij}}^{EB} = \sum_{k \in S_{ij}} a_{ijk}(\widehat{B_{ijk}}\ \widehat{\pi_{ijk}}^{D} + (1-\widehat{B_{ijk}})\widehat{\mu_{ik}}) + \sum_{k \notin S_{ij}} a_{ijk}\widehat{\mu_{ik}} , \quad (2.79)$$

$$\text{단, } \widehat{B_{ijk}} = \frac{d\ \widehat{\mu_{ik}}^{2}}{d\ \widehat{\mu_{ik}}^{2} + c_{ijk}(\widehat{\mu_{ik}} - (d+1)\ \widehat{\mu_{ik}}^{2})}$$

가중치 또는 축소인자(shrinkage factors)로 불리우는 $\widehat{B_{ijk}}$는 $\widehat{\pi_{ijk}}$의 분산에 대한 π_{ijk}의 분산의 비이고, $\widehat{\mu_{ik}}$는 $\widehat{\pi_{ik}}^{D}$로 대체되어 구해진다.

베이즈 추정량의 MSE는 $MSE(\widehat{\pi_{ij}}^{B}) = E(\widehat{\pi_{ij}}^{B} - \pi_{ij})^2$로 구해지고 i)과 ii)의 가정하에서 다음과 같이 계산된다.

$$\begin{aligned} MSE(\widehat{\pi_{ij}}^{B}) &= Var(\widehat{\pi_{ij}}^{B} - \pi_{ij}) \\ &= Var(\widehat{\pi_{ij}}^{B}) + Var(\pi_{ij}) - 2\,Cov(\widehat{\pi_{ij}}^{B}, \pi_{ij}) \\ &= Var(\pi_{ij}) - Var(\widehat{\pi_{ij}}^{B}) \\ &= d\left\{ \sum_{k \in S_{ij}} a_{ijk}^{2}(1-B_{ijk})\,\mu_{ik}^{2} + \sum_{k \notin S_{ij}} a_{ijk}^{2}\,\mu_{ik}^{2} \right\} \end{aligned}$$

경험적 베이즈 추정량의 MSE는 다음식으로 주어진다.

$$MSE(\widehat{\pi_{ij}}^{EB}) = MSE(\widehat{\pi_{ij}}^{B}) + E(\widehat{\pi_{ij}}^{EB} - \widehat{\pi_{ij}}^{B})^2 \qquad (2.80)$$

(2.80)식은 미지인 모수 μ_{ik}를 포함하고 있기 때문에 추정되어야 한다.

첫 번째 항 $MSE(\widehat{\pi_{ij}}^{B})$는 다음 식으로 추정된다(Jiang et al.(1998)).

$$mse_J(\widehat{\pi_{ij}}^{B}) = mse(\widehat{\pi_{ij}}^{B}) - \frac{J_i - 1}{J_i} \sum_{u=1}^{J_i} \left\{ mse_{(-u)}(\widehat{\pi_{ij}}^{B}) - mse(\widehat{\pi_{ij}}^{B}) \right\},$$

단, $mse(\widehat{\pi_{ij}}^{B}) = d\left\{ \sum_{k \in S_{ij}} a_{ijk}^{2}(1-B_{ijk})\,\mu_{ik}^{2} + \sum_{k \notin S_{ij}} a_{ijk}^{2}\,\widehat{\mu_{ik}}^{2} \right\}$,

$$mse_{(-u)}(\widehat{\pi_{ij}}^{B}) = d\left\{ \sum_{k \in S_{ij}} a_{ijk}^{2}(1-\widehat{B_{ijk(-u)}})\,\widehat{\mu_{ik(-u)}}^{2} + \sum_{k \notin S_{ij}} a_{ijk}^{2}\,\widehat{\mu_{ik(-u)}}^{2} \right\}$$

단, $\widehat{\mu_{ik(-u)}} = \dfrac{\sum_{j \neq u}^{J_i} \sum_{l=1}^{n_{ijl}} w_{ijkl}\, y_{ijkl}}{\sum_{j \neq u}^{J_i} \sum_{l=1}^{n_{ijl}} w_{ijkl}}$,

$$\widehat{B_{ijk(-u)}} = \frac{d\,\widehat{\mu_{ik(-u)}}^{2}}{d\,\widehat{\mu_{ik(-u)}}^{2} + c_{ijk}\left\{ \widehat{\mu_{ik(-u)}} - (d+1)\,\widehat{\mu_{ik(-u)}}^{2} \right\}}.$$

두 번째 항 $E(\widehat{\pi_{ij}}^{EB} - \widehat{\pi_{ij}}^{B})^2$ 은 다음의 Jackknife 추정법으로 추정될

수 있다(Shao and Tu(1995)).

$$E_J(\widehat{\pi}_{ij}^{EB} - \widehat{\pi}_{ij}^{B})^2 = \frac{J_i - 1}{J_i} \sum_{u=1}^{J_i} (\widehat{\pi}_{ij(-u)}^{EB} - \widehat{\pi}_{ij}^{EB})^2 ,$$

여기에서 $\widehat{\pi}_{ij(-u)}^{EB} = \sum_{k \in S_{ij}} a_{ijk} \{ \widehat{B}_{ijk(-u)} \widehat{\pi}_{ijk}^{D} + (1 - \widehat{B}_{ijk(-u)}) \widehat{\mu}_{ik(-u)} \}$

$$+ \sum_{k \notin S_{ij}} a_{ijk} \widehat{\mu}_{ik(-u)}$$

따라서 $MSE(\widehat{\pi}_{ij}^{EB})$ 의 추정량은 다음식으로 주어진다.

$$mse(\widehat{\pi}_{ij}^{EB}) = mse_J(\widehat{\pi}_{ij}^{B}) + E_J(\widehat{\pi}_{ij}^{EB} - \widehat{\pi}_{ij}^{B})^2 \qquad (2.81)$$

(5) 추정 결과

<표 2.6> 40개의 county에 대한 알콜중독률의 5개 추정량(%)

(Est.se=직접추정치의 표준오차, $\sqrt{Est.\ mse}$=EB추정치의 추정된 MSE의 제곱근)

	Estimator								
	Direct		Synthetic1	Synthetic2	Composite	Empirical Bayes		Sample Size	Number of Groups Observed in County
county	$\hat{\pi}_{ij}^{D}$	(Est.se)	$\hat{\pi}_{ij}^{S1}$	$\hat{\pi}_{ij}^{S2}$	$\hat{\pi}_{ij}^{C}$	$\hat{\pi}_{ij}^{EB}$	$\sqrt{Est.\ mse}$		
1	1.7	(2.4)	3.4	1.6	0.9	1.6	(0.33)	30	8
2	4.4	(2.0)	3.8	1.8	7.2	2.1	(0.35)	111	8
3	0.0	(0.0)	3.6	3.3	0.0	3.0	(0.85)	36	8
4	0.0	(0.0)	3.3	5.6	1.6	5.3	(1.79)	6	5
5	9.4	(4.8)	3.3	5.6	14.1	6.9	(1.78)	37	8
6	1.6	(1.1)	3.4	3.0	1.7	2.7	(0.67)	136	8
7	9.3	(5.8)	3.4	3.1	9.9	3.1	(0.81)	25	6
8	0.0	(0.0)	3.6	3.2	0.4	3.1	(0.84)	20	7
9	0.0	(0.0)	3.4	5.8	5.6	5.8	(1.93)	3	3
10	1.5	(1.3)	3.4	2.1	0.7	1.9	(0.54)	81	8
11	0.0	(0.0)	3.3	1.6	0.0	1.5	(0.33)	58	8
12	7.0	(6.8)	3.5	1.7	5.0	1.8	(0.35)	14	6
13	5.7	(3.8)	3.3	5.5	12.9	6.4	(1.75)	37	8
14	0.0	(0.0)	3.5	1.7	0.8	1.6	(0.33)	12	4
15	2.4	(1.4)	3.3	5.6	2.0	4.4	(1.56)	120	8
16	4.1	(3.5)	3.3	3.0	2.5	3.0	(0.77)	32	7
17	2.8	(2.4)	3.8	1.8	1.3	1.8	(0.37)	48	8
18	3.9	(1.1)	3.4	3.0	3.2	3.2	(0.60)	316	8
19	0.0	(0.0)	3.4	5.7	3.7	5.7	(1.95)	19	5
20	3.1	(3.9)	3.6	3.2	14.9	3.2	(0.82)	20	6
21	2.7	(1.6)	3.3	5.6	4.1	5.8	(1.50)	102	8
22	4.2	(1.8)	3.3	2.1	1.8	2.2	(0.42)	124	8
23	9.7	(2.7)	4.3	8.0	11.8	8.8	(2.11)	121	8
24	0.0	(0.0)	3.3	2.0	0.2	1.9	(0.54)	22	6
25	7.8	(4.7)	3.3	1.6	2.8	1.8	(0.33)	32	6
26	0.0	(0.0)	3.5	1.7	0.0	1.6	(0.37)	28	7
27	2.2	(1.8)	3.2	5.6	1.6	4.9	(1.74)	63	8
28	10.5	(13.7)	3.4	1.6	14.2	1.7	(0.35)	5	5
29	0.0	(0.0)	3.5	3.1	1.8	3.0	(0.81)	12	5
30	0.0	(0.0)	3.2	1.5	0.0	1.5	(0.33)	11	6
31	4.6	(3.2)	3.5	5.9	17.0	5.8	(1.87)	44	8
32	8.4	(3.8)	3.7	3.4	8.4	4.1	(0.84)	52	8
33	2.5	(1.3)	3.4	2.2	2.5	2.1	(0.50)	144	8
34	2.9	(2.4)	3.6	1.7	1.3	1.7	(0.35)	49	7
35	0.0	(0.0)	3.3	3.0	0.0	2.8	(0.77)	22	8
36	0.0	(0.0)	3.4	3.1	0.3	2.9	(0.82)	17	6
37	4.2	(4.0)	3.0	2.0	3.4	2.1	(0.54)	26	6
38	0.0	(0.0)	3.4	5.8	3.7	5.7	(1.97)	16	6
39	0.0	(0.0)	3.5	3.1	0.6	3.0	(0.81)	10	6
40	5.3	(1.9)	3.4	3.1	2.9	3.5	(0.69)	144	8

직접추정값은 변화가 심하고 경우에 따라서는 0으로 추정되며, 표준오차는

$\sqrt{\frac{\hat{\pi}_{ij}^{D}(1-\hat{\pi}_{ij}^{D})}{n_{ij}}}$로 추정될 수 있으며 또한 0으로 추정되는 경우가 있다. 합성추정값(S1)이 가장 안정적이며, 0으로 추정되는 값은 없고 약간의 변화성만 보인다. 반면, 복합추정값은 다른 추정값에 비해 훨씬 변화가 심하게 나타난다. 경험적인 베이즈 추정값들은 합성추정값(S2)과 매우 유사하며 추정된 MSE의 제곱근 값은 비교적 안정적으로 나타난다 (경험적 베이즈 추정량의 d 값은 $\frac{1}{3}$로 하였음). 특히 경험적 베이즈 추정량은 county의 표본크기가 작을 경우 매우 효과적임을 확인할 수 있다

<표 2.7>은 <표 2.6>에 나타나있는 대상 주의 모든 county들에 대한 알콜 중독률에 대한 추정결과이다.

<표 2.7> 40개의 county의 알콜중독률에 대한 5개 추정량들의 요약(%)

Estimator	Min.	Q1	Q2	Q3	Max.	Mean	S.D
Direct	0.0	0.0	2.2	4.3	10.5	2.8	3.2
Synthetic 1	3.0	3.3	3.4	3.5	4.3	3.5	0.2
Synthetic 2	1.5	1.8	3.0	4.4	8.0	3.2	1.7
Composite	0.0	0.4	1.7	4.6	17.5	3.7	4.8
EB	1.5	1.8	2.8	4.2	8.8	3.2	4.8

여기에서 합성추정값들과 복합추정값들의 평균이 직접추정값들의 평균보다 높게 나타나는데 이는 합성추정값과 복합추정값들은 직접추정값들에 비해 0으로 추정된 값들이 거의 없기 때문이다.

2.4.4 MCMC기법을 이용한 소지역 추정사례

(1) 서 론

소지역 추정문제를 해결하기 위한 다양한 방법의 연구가 최근에 들어 많은 사람들의 관심에 의해 진행되고 있는데, 특히 베이즈 방법은 모형을 통해 소지역들을 체계적으로 연결하여 소지역 추론을 할 수 있다는 점에서 광범위하게 이용되고 있다. 계층적 베이즈(HB) 방법과 경험적 베이즈(EB) 방법에 관한 응용 및 일반 이론에 관한 내용은 Datta and Ghosh(1991), Fay and Herriot(1979), Ghosh and Lahiri(1987, 1992), Prasad and Rao(1990), Stroud(1987, 1991) 등에 의해 연구되었으며, 주로 연속형 자료를 갖는 변량들에 관한 내용을 다루었다. 그러나 많은 경우 조사자료들은 이산형이거나 범주형 자료일 수 있으며, 이러한 경우에 위의 방법들을 직접적으로 이용하는 것은 한계가 있다.

이진 조사자료(binary survey data)로 경험적 베이즈(EB) 방법 또는 계층적 베이즈(HB) 방법을 이용하여 소지역의 비율을 추정하는 방법들은 Dempster and Tomberlin(1980), MacGibbon and Tomberlin(1989), Malec, Sedransk, and Tompkins(1993) 등에 의해 연구되었다. 또한, Nandram and Sedransk(1993)은 이단계 집락표본(two-stage cluster sample)의 이진 자료에 대한 베이즈 추정방법을 소개하였고, 뒤이어 Stroud(1994)에 의해 이단계 표본추출뿐만 아니라 단순임의추출, 층화추출, 집락추출된 조사자료에 대한 베이즈 추정연구가 소개되었다.

이진 모형(binary model)들은 이산형 자료와 연속형 자료를 동시에 통합하는 일반화된 선형모형(Generalized Linear Model)의 일부분으로 구분된다. Ghosh, Natarajan, Stroud, and Carlin(1998)은 이산형 또는 범주형으로 구분되는 조사자료(Survey data)를 계층모형에 적합시켜 베이즈 방법으로 소지역 추정을 해결하는 방법을 소개하였다.

베이즈 방법을 이용한 소지역 추정에는 주로 MCMC(Markov Chain Monte Carlo) 적분기법이 적용된다. 깁스 샘플러는 이러한 MCMC 방법의 일종이

다. 여기에 소개되는 자료는 1991년 캐나다 내의 지리적으로 구분된 15개 지역의 표본에 대해서 "당신이 근무하고 있는 작업장에서 건강상의 유해요소에 노출되어 나쁜 영향을 받은 적이 있는가?"라는 질문에 대한 응답 결과 (① 그런적이 있다(yes) ② 그런적이 없다(no) ③ 건강상의 유해요소에 노출된 적이 없다(not exposed) ④ 해당 사항이 없다(not applicable or not stated))에 관한 조사자료이며, 이 조사자료들은 각각의 지역에 대해서 특성화 기준(연령별: 40세 이하, 41세 이상, 성별: 남, 여)에 의해 범주화 되었다. 이 조사자료를 기반으로 각각의 지역내에서 연령별, 성별 기준에 따라 질문문항에 대한 응답비율을 추정하고자 하였다.

(2) 계층 모형

I개의 소지역이 있을 때, Y_{ij}를 i번째 소지역 내에 있는 j번째 단위에 대한 최소충분통계량이라 하자($i=1,2,\cdots,I$; $j=1,2,\cdots,n_i$). 이때 Y_{ij}는 서로 독립이며 다음과 같은 확률밀도함수를 갖는다고 가정한다.

$$f(y_{ij}|\theta_{ij},\phi_{ij}) = \exp\left\{\frac{y_{ij}\theta_{ij} - \psi(\theta_{ij})}{\phi_{ij}} + \rho(y_{ij};\phi_{ij})\right\} \qquad (2.82)$$

즉, Y_{ij}의 확률밀도함수를 지수족(Exponential Family)으로 가정하며, 여기에서 θ_{ij}는 정준모수(canonical parameter), ϕ_{ij} (>0)는 산포모수(dispersion parameter)를 나타내며, 산포모수 ϕ_{ij}는 기지로 가정한다. 통상적으로 (5.51)의 가정하에서는 $g(\theta_{ij}) = \psi'(\theta_{ij}) \quad = E(Y_{ij}|\theta_{ij})$ ($\equiv \mu_{ij}$)인 관계가 성립한다.

자연모수(natural parameter) θ_{ij}는 소지역간의 상관성을 고려하여 다음과 같은 구조로 모형화한다.

$$h(\theta_{ij}) = \quad \boldsymbol{x}_{ik}^{T}\boldsymbol{\beta} + u_i + \varepsilon_{ij}\ ,\ i=1,2,\cdots,I\ ,\ j=1,2,\cdots,n_i \qquad (2.83)$$

여기에서 h는 엄밀증가함수(strictly increasing function)이고, x_{ik}는 기지인 $p\times 1$ 계획벡터(design vector)이며, $\beta=(\beta_1,\beta_2,\cdots,\beta_p)^T$는 $p\times 1$ 회귀계수(random regression coefficient) 벡터, u_i는 랜덤효과들이다. 오차 항 ε_{ij}는 $\varepsilon_{ij}\sim iid\ N(0,\sigma^2)$이고, 랜덤 항 u_i는 $u_i\sim iid\ N(0,\sigma_u^2)$이며, ε_{ij}와 u_i는 서로 독립이라 가정한다.

위의 (2.82)을 계층적 구조를 갖는 구조화된 모형으로 다음과 같이 표현할 수 있다. 아래의 표현식에서

$\theta=(\theta_{11},\theta_{12},\cdots,\theta_{1n_1},\cdots,\theta_{I1},\theta_{I2},\cdots,\theta_{In_I})^T$, $R_u=\sigma_u^{-2}$, $R=\sigma^{-2}$을 나타낸다.

(Ⅰ) $Y_{ij}|\theta,\beta,u,R_u=r_u,R=r$

$$\overset{\text{ind}}{\sim}\ f(y_{ij}|\theta_{ij},\phi_{ij})=\exp\left\{\frac{y_{ij}\theta_{ij}-\phi(\theta_{ij})}{\phi_{ij}}+\rho(y_{ij};\phi_{ij})\right\}$$

(Ⅱ) $h(\theta_{ij})|\beta,u,R_u=r_u,R=r$

$$\overset{\text{ind}}{\sim}\ N(x_{ij}^T\beta+u_i\ ,\ r^{-1})$$

(Ⅲ) $u_i|\beta,R_u=r_u,R=r$

$$\overset{\text{ind}}{\sim}\ N(0\ ,\ r_u^{-1})$$

여기에서 모수 β, $R_u=r_u$, $R=r$ 의 사전분포들은 다음과 같이 가정한다

(Ⅳ) $\beta\sim uniform(R^p)$, $(p<m)$,

$R_u\sim \text{gamma}(\frac{a}{2}\ ,\ \frac{b}{2})$,

$R\sim \text{gamma}(\frac{c}{2}\ ,\ \frac{d}{2})$ 이고, β , R_u, R은 서로 독립

(단, $f(z) \propto \exp(-az)\, z^{\beta-1} I_{(0,\infty)}(z)$일 때, $Z \sim \text{gamma}(\alpha, \beta)$)

위에서 언급한 $g(\theta_{ij})$는 Y_{ij}의 조건부 분포의 중심부분으로써 우리의 관심을 여기에 집중시킨다. 즉,

$\boldsymbol{y}=(y_{11}, y_{12}, \cdots, y_{1n_1}, \cdots, y_{I1}, y_{I2}, \cdots, y_{In_I})^T$가 주어진 상태에서 $g(\theta_{ij})$들의 평균, 분산, 공분산 계산에 관심이 있다.

사전분포 (Ⅳ)를 가정한 모형 (2.83)에서 $\boldsymbol{y}$ 에 대한 θ_{ij}의 조건부 결합 사후확률분포는 $a>0$, $c>0$, $\sum_{i}^{m} n_I - p + d > 0$, $m+b>0$ 일 때, 모든 y_{ij}와 $\phi_{ij}(>0)$에 대해서 $\int_{\underline{\theta}_{ij}}^{\bar{\theta}_{ij}} \exp\left\{\frac{\theta y_{ij}-\phi(\theta)}{\phi_{ij}}\right\} h'(\theta)\, d\theta < \infty$ 이고, 주어진 $\boldsymbol{y}$에 대한 θ_{ij}들의 결합 사후확률분포는 진분포(proper distribution)이다. 여기에서 θ_{ij}의 적분구간은 $\pm\infty$를 포함하는 개구간(open interval)이다.

일반화된 선형모형(GLM)에서 $Y_{ij}|p_{ij} \sim Bin(n_{ij}, p_{ij})$인 경우를 생각해 보자. 이때 $Y_{ij}|p_{ij} \propto \exp\left\{y_{ij} \log\frac{p_{ij}}{1-p_{ij}} + n_{ij}\log(1-p_{ij})\right\}$이며, 정준연결(canonical link)을 고려하면, $\theta_{ij} = \log(p_{ij}/1-p_{ij})$ 이고, $g(\theta_{ij}) = \phi'(\theta_{ij})/n_{ij} = \exp\{\theta_{ij}/(1+\exp(\theta_{ij}))\}$이 된다. $g(\theta_{ij})$의 평균, 분산, 공분산 등은 θ_{ij}의 분포로부터 계산되며, 이를 위한 조건부 사후확률분포들은 위의 (Ⅰ)~(Ⅳ)를 이용하여 계산하면 다음의 (a)~(e)와 같이 주어진다.

(a) $\boldsymbol{\beta}|\,\boldsymbol{\theta}, \boldsymbol{u}, r_u, r, \boldsymbol{y}$

$$\sim N_p\left(\frac{\sum_i\sum_j [\,h(\theta_{ij})\,\boldsymbol{x}_{ij} - \boldsymbol{x}_{ij}\,u_i\,]}{\sum_i\sum_j \boldsymbol{x}_{ij}\,\boldsymbol{x}_{ij}^T},\quad r^{-1}\left(\sum_i\sum_j \boldsymbol{x}_{ij}\,\boldsymbol{x}_{ij}^T\right)^{-1}\right)$$

(b) $u_i \,|\, \boldsymbol{\theta}, \boldsymbol{\beta}, r_u, r, \boldsymbol{y}$

$$\overset{ind}{\sim} N(rn_i + r_u)^{-1} r \sum_j [h(\theta_{ij}) - x_{ij}^T \beta], \quad (rn_i + r_u)^{-1})$$

(c) $R \mid \theta, \beta, u, r_u, y$

$$\sim G\left(\frac{1}{2}\left\{c + \sum_i \sum_j [h(\theta_{ij}) - x_{ij}^T \beta - u_i]^2\right\}, \frac{1}{2}(d + \sum_{i=1}^{I} n_i)\right)$$

(d) $R_u \mid \theta, \beta, u, r, y$

$$\sim G\left(\frac{1}{2}(a + \sum_{i=1}^{I} u_i^2), \frac{1}{2}(b + \sum_{i=1}^{I} n_i)\right),$$

(e) $\theta_{ij} \mid \beta, u, r_u, r, y$

$$\overset{ind}{\sim} \pi(\theta_{ij} \mid \beta, u, r_u, r, y)$$

$$\propto \exp\left\{\frac{y_{ij}\theta_{ij} - \phi(\theta_{ij})}{\phi_{ij}} - \frac{r}{2}[h(\theta_{ij}) - x_{ij}^T \beta - u_i]^2\right\} h'(\theta_{ij})$$

$E(\theta_{ij} \mid y)$, $V(\theta_{ij} \mid y)$, $cov(\theta_{ij}, \theta_{i'j'})$ $(i, j) \neq (i', j')$등의 θ에 관한 추론은 MCMC(Monte Carlo Markov Chain) 방법을 이용할 수 있으며, 깁스 샘플링(Gibbs sampling)은 이러한 MCMC 방법의 일종이다.

위의 (a)~(e)의 조건부 분포를 이용한 깁스 알고리즘은 다음의 절차에 의해 이루어진다.

(i) $\theta_{ij} = \theta_{ij}^{(0)}$, $u_i = u_i^{(0)}$, $r = r^{(0)}$를 초기값으로 하여 위의 (a)로부터 β를 생성하고 이것을 $\beta^{(1)}$ 이라하자.

(ii) $\beta = \beta^{(1)}$, $\theta_{ij} = \theta_{ij}^{(0)}$, $r = r^{(0)}$, $r_u = r_u^{(0)}$ 를 이용하여 (b)로부터 $u_i = u_i^{(1)}$을 생성한다.

(iii) $\beta = \beta^{(1)}$, $u_i = u_i^{(1)}$, $\theta_{ij} = \theta_{ij}^{(0)}$를 이용하여 (c)로부터 $r = r^{(1)}$을 생성한다.

(iv) $u_i = u_i^{(1)}$을 이용하여 (d)로부터 $r_u = r_u^{(1)}$ 을 생성한다.

(v) $\beta = \beta^{(1)}$, $u_i = u_i^{(1)}$, $r = r^{(1)}$, $\theta_{ij} = \theta_{ij}^{(0)}$를 이용하여 (e)로부터 $\theta_{ij} = \theta_{ij}^{(1)}$을 생성한다($i = 1, 2, \cdots, m$).

(vi) 절차 (i)~(v)를 한 사이클로 하여 같은 과정을 반복 수행한다.

수렴이 이루어지는 시점 t까지 충분히 반복한 후, 이 후부터 얻어지는 J개의 표본 $\{ \beta^{(t+k)}, u_1^{(t+k)}, u_2^{(t+k)}, \cdots, u_I^{(t+k)}, r^{(t+k)}, r_u^{(t+k)}, \theta_{1j}^{(t+k)}, \theta_{2j}^{(t+k)}, \cdots, \theta_{Ij}^{(t+k)}; k = 1, 2, \cdots, J\}$을 $\beta, u_1, u_2, \cdots, u_I, r, r_u, \theta_{1j}, \theta_{2j}, \cdots, \theta_{Ij}$의 결합 사후분포로부터 얻은 표본으로 간주하며, θ_{ij} 의 사후평균, 사후분산은 위에서 생성된 J개의 표본을 이용하여 추정한다.

I개의 소지역이 주어진 경우, 각 소지역내에서 여러개의 단위가 선택되고, 선택된 단위들의 응답은 서로 독립적이며 J개의 범주로 분류될 수 있다고 할 때, p_{ijk}를 i번째 소지역에서 k번째로 선택된 단위가 j번째 범주일 확률로 정의하자($i = 1, 2, \cdots, I$; $j = 1, 2, \cdots, J$; $k = 1, 2, \cdots, n_i$). 이 때 i번째 소지역 내에서 k번째로 선택된 단위 $Z_{ijk}(j = 1, 2, \cdots, J)$는 J개의 범주 중 하나로 구분될 수 있고, 선택된 단위의 응답은 서로 독립인 multinomial($\sum_j Z_{ijk}; p_{i1k}, p_{i2k}, \cdots, p_{iJk}$)분포를 따른다고 가정할 수 있다.

만약 $Y_{ijk}(j = 1, 2, \cdots, J)$가 서로 독립인 $poisson(\zeta_{ijk})$ 분포를 따르고, $p_{ijk} = \frac{\zeta_{ijk}}{\sum_{j=1}^{J} \zeta_{ijk}}$ ($j = 1, 2, \cdots, J$)일 경우, multinomial 분포와 $poisson$분포간에는 $(Z_{i1k}, Z_{i2k}, \cdots, Z_{iJk}) \overset{d}{=} (Y_{i1k}, Y_{i2k}, \cdots, Y_{iJk}) | \sum_j Y_{ijk} = t_{ik}$ 의 관계가 성립한다.

θ_{ijk}를 $\theta_{ijk} = \log \zeta_{ijk}$로 놓았을 때 $\zeta_{ijk} = \exp(\theta_{ijk})$이므로 $p_{ijk} = \dfrac{\exp(\theta_{ijk})}{\sum_{j=1}^{J} \exp(\theta_{ijk})}$로 표현될 수 있다. 여기에서의 관심은 p_{ijk}에 관한 추론이며, 이에 앞서 p_{ijk}의 추론에 필요한 θ_{ijk}의 모형을 다음과 같은 모형으로 구조화할 수 있다.

$$h(\theta_{ijk}) = \boldsymbol{x}_{ijk}^{T}\boldsymbol{\beta} + u_{ij} + \varepsilon_{ijk} \qquad (2.84)$$

$$(i = 1, 2, \cdots, I;\ j = 1, 2, \cdots, J;\ k = 1, 2, \cdots, n_i)$$

여기서 $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$는 $p \times 1$벡터, $\boldsymbol{x}_{ijk} = (x_{ijk1}, x_{ijk2}, \cdots, x_{ijkp})^T$는 $p \times 1$ 계획벡터이다. 오차 항 ε_{ijk}는 $\varepsilon_{ijk} \sim iid\ N(0, \sigma^2)$이고, 랜덤 항 u_{ij}는 $u_{ij} \sim iid\ N(0, \sigma_u^{\,2})$이며, ε_{ijk}와 u_{ij}는 서로 독립이라 가정한다.

식(2.82)과 (2.84)을 계층적인 구조의 모형으로 다음과 같이 구조화한다. 다음의 표현에서 $\boldsymbol{\theta} = (\theta_{111}, \theta_{112}, \cdots, \theta_{11n_1}, \cdots, \theta_{IJ1}, \theta_{IJ2}, \cdots, \theta_{IJn_I})^T$, $R = \sigma^{-2}$, $R_u = \sigma_u^{\,-2}$, $\boldsymbol{u} = (u_{11}, u_{12}, \cdots, u_{1J}, \cdots, u_{m1}, u_{m2}, \cdots, u_{mJ})^T$를 나타낸다.

(Ⅰ) $Y_{ijk} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{u}, R_u = r_u, R = r$

$$\overset{\text{ind}}{\sim}\ f(y_{ijk} \mid \theta_{ijk}, \phi_{ijk}) = \exp\left\{\frac{y_{ijk}\theta_{ijk} - \phi(\theta_{ijk})}{\phi_{ijk}} + \rho(y_{ijk}; \phi_{ijk})\right\}$$

(Ⅱ) $h(\theta_{ijk}) \mid \boldsymbol{\beta}, \boldsymbol{u}, R_u = r_u, R = r$

$$\overset{\text{ind}}{\sim}\ N(\boldsymbol{x}_{ijk}^{T}\boldsymbol{\beta} + u_{ij},\ r^{-1})$$

(Ⅲ) $u_{ij} \mid \boldsymbol{\beta}, R_u = r_u, R = r$

$$\overset{\text{ind}}{\sim}\ N(0,\ r_u^{-1})$$

여기에서 모수 β, $R_u = r_u$, $R = r$ 의 사전분포를 다음과 같이 가정한다

(Ⅳ) $\beta \sim uniform(R^p)$, $(p < m)$,

$$R_u \overset{\text{ind}}{\sim} G\left(\frac{a}{2}, \frac{b}{2}\right),$$

$R \sim \text{gamma}\left(\frac{c}{2}, \frac{d}{2}\right)$ 이고, β, R_u, R은 서로 독립

$p_{ijk} = \dfrac{\exp(\theta_{ijk})}{\sum_{j=1}^{J} \exp(\theta_{ijk})}$ 의 사후평균, 분산, 공분산을 계산하기 위해서 필요한 사후분포들은 다음과 같다.

(a) $\beta \mid \theta, u, r_u, r, y$

$$\sim N_p\left(\frac{\sum_{i,j,k}[h(\theta_{ijk})\, x_{ijk} - u_{ij}]}{\sum_{i,j,k} x_{ijk}\, x_{ijk}^T}, \; r^{-1}\left(\sum_{i,j,k} x_{ijk}\, x_{ijk}^T\right)^{-1}\right)$$

(b) $u_{ij} \mid \theta, \beta, r_u, r, y$

$$\overset{\text{ind}}{\sim} N\left((rn_i + r_u)^{-1} r \sum_j (h(\theta_{ijk}) - x_{ijk}^T \beta), \; (rn_i + r_u)^{-1}\right)$$

(c) $R \mid \theta, \beta, u, r_u, y$

$$\sim \text{gamma}\left(\frac{1}{2}\left\{c + \sum_{i,j,k}[h(\theta_{ijk}) - x_{ijk}^T \beta - u_{ij}]^2\right\}, \; \frac{1}{2}(d + J\textstyle\sum_i n_i)\right)$$

(d) $R_u \mid \theta, \beta, u, r, y$

$$\sim G\left(\frac{1}{2}(a + \sum_i \sum_j u_{ij}^2), \; \frac{1}{2}(b + IJ)\right), \quad s = 1, 2, \cdots, p$$

(e) $\theta_{ijk} \mid \theta, \beta, u, r_u, r, y$

$$\overset{\text{ind}}{\sim} \pi(\theta_{ijk} \mid \theta, \beta, u, r_u, r, y)$$

$$\propto \exp\left\{\frac{y_{ijk}\theta_{ijk}-\phi(\theta_{ijk})}{\phi_{ijk}} - \frac{r}{2}[h(\theta_{ijk}) - \boldsymbol{x}_{ijk}^{T}\boldsymbol{\beta} - u_{ij}]^2\right\} h'(\theta_{ijk})$$

(3) 추정 결과

모형 (2.84)에서 k를 연령과 성별을 나타내는 (a, s)로 구분하여 표기하도록 하고, 회귀방정식 $\boldsymbol{x}_{ijk}\boldsymbol{\beta} = \mu + \tau_a^A + \tau_s^S + \tau_j^J + \tau_{as}^{AS} + \tau_{aj}^{AJ} + \tau_{sj}^{SJ}$ 을 적합시켜 소지역에서의 특성별 추정값들을 계산하고자 한다. 여기에서 μ는 일반효과, τ_a^A는 a번째 연령그룹에서의 주효과, τ_s^S는 s번째 성별그룹에서의 주효과, τ_j^J는 j번째 응답범주와 관련된 주효과, τ_{as}^{AS}는 a번째 연령그룹과 s번째 성별그룹의 교호작용의 효과, τ_{aj}^{AJ}는 a번째 연령그룹과 j번째 응답범주의 교호작용의 효과, τ_{sj}^{SJ}는 s번째 성별그룹과 j번째 응답범주의 교호작용의 효과를 나타낸다. 모든 a, s, j에 대하여 $\tau_1^A = \tau_1^S = \tau_1^J = \tau_{a1}^{AS} = \tau_{1s}^{AS} = \tau_{a1}^{AJ} = \tau_{1j}^{AJ} = \tau_{s1}^{SJ} = \tau_{1j}^{SJ} = 0$ 의 제한조건을 가정한다.

가정된 모형을 이용한 소지역 특성별 추정결과는 다음의 <표 2.8>과 같다. 결과적으로 경험적 베이즈 추정량의 표준오차가 표본 추정량의 표준오차에 비해 상대적으로 매우 작은 사실을 확인할 수 있다.

<표 2.8> Impact of Exposure to Health Hazards in the Workplace

Category	Response	Sample		H. Bayes	
		Proportions	SD	Proportions	SD
Region=2	Total n=294				
M, Age<40	Yes	.400	.100	.373	.042
	No	.383	.101	.345	.041
	Not exposed	.150	.119	.199	.031
	NA/NS	.067	.125	.083	.015
F, Age<40	Yes	.257	.100	.266	.035
	No	.284	.098	.279	.035
	Not exposed	.311	.097	.274	.036
	NA/NS	.148	.107	.181	.026
M, Age≥40	Yes	.111	.111	.184	.028
	No	.153	.109	.176	.027
	Not exposed	.167	.108	.156	.026
	NA/NS	.569	.077	.484	.040
F, Age≥40	Yes	.159	.098	.110	.019
	No	.091	.102	.103	.018
	Not exposed	.125	.010	.134	.022
	NA/NS	.625	.065	.654	.034

Category	Response	Sample		H. Bayes	
		Proportions	SD	Proportions	SD
Region=3	Total n=740				
M, Age<40	Yes	.294	.070	.311	.029
	No	.426	.063	.395	.032
	Not exposed	.203	.075	.186	.023
	NA/NS	.077	.080	.108	.015
F, Age<40	Yes	.246	.064	.235	.024
	No	.273	.063	.287	.026
	Not exposed	.180	.067	.204	.023
	NA/NS	.301	.062	.274	.026
M, Age≥40	Yes	.156	.069	.154	.019
	No	.150	.069	.165	.020
	Not exposed	.100	.071	.112	.016
	NA/NS	.594	.048	.569	.028
F, Age≥40	Yes	.064	.063	.071	.010
	No	.086	.063	.091	.012
	Not exposed	.111	.062	.099	.013
	NA/NS	.739	.033	.739	.021

Category	Response	Sample		H. Bayes	
		Proportions	SD	Proportions	SD
Region=8	Total n=1707				
M, Age<40	Yes	.274	.047	.279	.021
	No	.360	.044	.362	.023
	Not exposed	.253	.048	.253	.020
	NA/NS	.113	.052	.106	.012
F, Age<40	Yes	.199	.042	.196	.016
	No	.267	.040	.275	.019
	Not exposed	.289	.040	.297	.019
	NA/NS	.245	.041	.234	.017
M, Age≥40	Yes	.113	.047	.130	.013
	No	.166	.046	.174	.016
	Not exposed	.217	.044	.195	.017
	NA/NS	.504	.035	.501	.022
F, Age≥40	Yes	.087	.042	.076	.009
	No	.123	.041	.110	.011
	Not exposed	.119	.041	.131	.012
	NA/NS	.671	.025	.683	.017

제3장 외국 소지역 실업통계 작성 사례

3.1 미 국

미국의 노동력 조사를 위한 표본관리체계는 4-8-4 연동교체체계로서 표본조사 가구로 한번 선정하면 4개월간 계속 조사를 실시하고 8개월간은 조사를 중지한 후, 다시 표본가구로 편입하여 4개월간 연속조사를 실시하는 제도이다. 이는 월간 노동력 변화와 연간 노동력 변화추세에 대한 추정값의 신뢰성을 높이고 표본조사 가구의 응답 부담을 줄이기 위한 제도이다. 먼저 주 단위의 실업통계 작성의 발전과정을 알아보고, 다음에는 주의 실업통계 작성 절차에 대해서 설명한다.

3.1.1 실업통계 발전과정

1950년대에 노동성의 고용 훈련 행정국에서 각 주 간에 실업률의 추정값을 비교할 수 있도록 실업자의 추정기법을 개발하여 책자로 발간하였다. 이 책자(handbook)를 근거로 하여 50년대 후반에는 대규모의 통계조사를 위한 비용 부담을 없애고, 소규모 조사를 통해서 연속적인 여러 단계를 거치는 주 단위 실업통계 작성기법을 "Handbook"방법으로 공식화하였다. 1950년대의 실업통계 작성은 주로 실업보험 신청자료를 이용하였다.

1972년에는 노동 통계국에서 주 단위의 이용 가능한 노동력, 취업과 실업의 추정에 대한 방법과 개념을 연구하기 시작하였고, 1973년에는 통계국이 주관이 되어, 경상인구조사(CPS : Current Population Survey)의 개념, 정의, 추정과 handbook방법을 결합하여 주 단위와 주의 세부단위까지의 노동력을 추정할 수 있는 기법을 개발하였다.

1976년 이후에는 모든 주 단위 실업통계의 추정값에 대한 신뢰도를 높이기 위해서 각 주별로 표본 가구 수를 몇 배씩 증가시켰으며, 이후부터 경상

인구조사의 자료를 이용한 노동력의 추정값을 공식적으로 발표하기 위해 실업률을 6%라고 가정했을 때, 실업통계에 대한 변동계수의 최대 예상 허용값을 10%로 하였다. 1978년부터 규모가 큰 10개주(California, Florida, Illinois, Massachusetts, Michigan, New Jersey, New York, Ohio, Pennsylvania, Texas)와 2개 대도시(Los Angels, New York City)의 노동력 통계는 CPS 자료에서 직접 추정법으로 계산한 결과를 공식 통계로 사용토록 하였다.

실업보험의 데이터베이스를 계속적으로 개선하였으며, 특히 1976~78년 사이에 실업보험 신청자료를 모든 주에 대해서 표준화하고, CPS조사의 조사주간(매월 12일을 포함한 주)에 실업자로 인정받는 실업 보험 신청자는 자동으로 데이터베이스에 등록되도록 하는 데이터 관리체계를 개발하여 CPS자료와 연계한 추정법을 개발하였다.

1985년에는 1980년 센서스 정보를 이용하여 state-based CPS 표본설계를 완성하였으며, North Carolina주를 CPS 자료에서 노동력 통계를 직접 추정하여 공표 하는 주로 포함시켰고, 대규모 11개 주(노동력 통계를 직접 추정법으로 추정하는 주)의 월별 실업통계에 대한 목표변동계수를 8%로 낮추었다. 또한 나머지 39개 주의 연평균 실업통계에 대한 목표변동계수를 8%로 정하였으며, 이 때 실업률의 참값은 6%인 것으로 가정하였다.

1989년까지는 직접 추정법을 적용하지 않은 39개 주와 substate의 실업수준의 공식적 월별 추정값은 Handbook방법을 적용하여 계산되었으나, 1989년 초부터 39개 주에 대한 노동력 통계 작성에서는 노동 통계국에서 개발한 시계열 모형 기법을 표준 추정법으로 채택하였다.

1994년에는 좀 더 고급회귀모형을 개발하여 소규모의 39개 주에 대한 실업통계 작성에 적용하였다.

1992년에는 주 단위의 추정값에 계절 조정을 적용하였고 1994년에는 1990년 센서스 자료를 이용하여 CPS를 재 설계하였으며, 또한 모든 조사에서 컴

퓨터 보조면접을 실시할 수 있도록 설문지를 재구성하였다. 새로 개편된 CPS의 표본설계는 1995년 중반까지 단계적으로 도입하였다.

1996년부터는 예산 절감으로 CPS표본 규모를 56,000가구에서 50,000가구로 축소했기 때문에 모든 주의 실업통계는 CPS자료에서 직접 추정하는 방식을 적용하지 않고 시계열 회귀모형을 적용하여 작성하도록 하였으며, 로스엔젤스와 뉴욕시의 실업통계 생산에 계절 조정 자료를 적용하였다. 그러나 소지역 단위인 노동시장단위(LMA : labor market area)의 실업통계를 생산하는데 좀 더 간편화된 회귀모형 같은 통계적 모형을 적용하기보다는 해당 LMA의 취업과 실업의 구조적 분석을 통한 building block절차를 적용하지만, 마지막 주정부내의 합계는 공식적인 주단위의 실업통계와 일치하도록 비례조정을 하였다.

3.1.2 회귀 모형 추정법

회귀모형을 이용한 추정법은 합성 추정값이 지역 변동을 충분히 설명하지 못한 점을 보완하고자 실직 보험 신청자료와 구직 등록 신청자료를 이용하여 1970년대 후반부터 시군구의 실업통계 작성에 이용되었던 기법이다. CPS 1차 추출단위(PSU : Primary Samplint Unit)의 추정치를 종속변수로 사용하고, 다음과 같은 몇 개의 적절한 독립변수를 선택한 회귀모형을 고려한다.

(1) 다음의 범주에서의 합성추정치

① Occupation-Sex-Race

② Marital Status-Age-Sex-Race

(2) 센서스에서 추정한 실업자 총계 대비 3-4월의 실직보험 가입자에 대한 가입비율(%)

(3) 연말 자료에서 공표되는 "70-step"의 실업자 추정치(실업률)

5개월 간의 월별 CPS 추정값들의 평균을 종속변수로 사용하기 위해,

CPS-PSU와 SMSA를 대응시켰을 때 150개의 SMSA 중 122개의 SMSA가 완전 대응되며 이러한 지역을 회귀모형 추정을 위한 자료로 활용하였다. 실업률 추정을 위해 적합한 회귀모형은 다음과 같다.

$$\hat{Y} = 0.008 - 0.201 X_1 + 0.680 X_2 + 0.404 X_3 , \qquad (3.1)$$

$$\text{Residual Mean Square} = 0.868 \times 10^{-4} ,$$

$$\text{추정치의 표준오차} = 0.932 \times 10^{-2} ,$$

$$R^2 = 0.546 ,$$

여기에서 Y = '5개월 간의 월별 CPS 실업추정치의 평균', X_1 = '전체 실업자 대비 보험가입 실업자의 비율(%)', X_2 = '최종 공표된 연말 실업률(final annual 70-step estimates)', X_3 = 'Marital Status-Age-Sex-Race 범주에서 계산된 실업률의 합성추정치'를 나타낸다.

위 회귀모형의 타당성 검증을 위하여 센서스 자료를 이용한 회귀모형의 적합결과는 다음과 같이 주어지며, 실업률 추정에서 회귀모형의 적용가능성을 보여준다.

$$\hat{U} = 0.009 + 0.012 X_1 + 0.586 X_2 + 0.540 X_3 , \qquad (3.2)$$

$$\text{Residual Mean Square} = 0.213 \times 10^{-4} ,$$

$$\text{추정치의 표준오차} = 0.461 \times 10^{-2} ,$$

$$R^2 = 0.859 ,$$

여기에서 U는 센서스 자료에서 계산한 실업률 추정치이다.

5개월 간의 월별 CPS 실업 추정치의 평균을 종속변수로 사용했을 때, 변동의 55%가 회귀모형으로 설명될 수 있고, 센서스의 실업률 추정치를 종속변수로 사용한 경우에는 변동의 86%가 회귀모형으로 설명될 수 있다. CPS 추정치의 표본변동이 커질수록 설명변수들의 기여도가 낮아진다는 것을 알 수 있다.

독립변수를 변화시켜 가면서 추가로 2종의 회귀모형을 추정해 보기로 한다. 먼저 독립변수로써 X_4를 CPS자료에 의한 추정치를 benchmarking하기 전의 실업 추정값으로 취하여 회귀모형을 적합했을 때 다음과 같은 회귀모형을 얻을 수 있다.

$$\hat{Y} = 0.012 - 0.252X_1 + 0.299X_3 + 0.078X_4 \quad , \qquad (3.3)$$

Residual Mean Square = 0.833×10^{-4} ,

추정치의 표준오차 = 0.912×10^{-2} ,

$R^2 = 0.565$

대응되는 회귀모형은 다음과 같다.

$$\hat{U} = -0.006 + 0.005X_1 + 0.477X_3 + 0.564X_4 \quad , \qquad (3.4)$$

Residual Mean Square = 0.228×10^{-4} ,

추정치의 표준오차 = 0.478×10^{-2} ,

$R^2 = 0.849$

식(3.3)의 회귀모형은 총변동의 약 57%를 설명하며, 회귀모형 (3.1)보다는 어느 정도 개선되었다.

두 번째 경우에는 최종 공표된 연말 실업률 X_2대신 월별 추정치의 11개월 간의 평균 X_5와 Occupation-Sex-Race의 3개의 범주에서 구한 합성추정치 X_6를 독립변수로 하였을 때 계산된 회귀모형의 적합 결과이다.

$$\hat{Y}'' = 0.009 - 0.210X_1 + 0.640X_5 + 0.444X_6 \quad , \qquad (3.5)$$

Residual Mean Square = 0.883×10^{-4} ,

추정치의 표준오차 = 0.939×10^{-2} ,

$R^2 = 0.539$

센서스 자료 추정치를 이용한 적합 결과는 다음과 같다.

$$\hat{U}'' = -0.008 - 0.011X_1 + 0.532X_5 + 0.617X_6 \quad , \qquad (3.6)$$

Residual Mean Square = 0.194×10^{-4} ,

추정치의 표준오차 = 0.440×10^{-2} ,

$R^2 = 0.872$

이상에서 살펴보았듯이 센서스 자료의 추정치를 종속변수로 사용했을 경우 회귀식은 약 85~87%의 설명력을 보이나, CPS자료 추정치를 사용했을 경우에는 표본 추출에 의한 변동에 기인하여 회귀식의 설명력은 약 54~56%의 범위 정도에 있음을 확인할 수 있다.

<표3.1>은 변수들 간의 상관관계를 나타낸 표이며, 70-step 추정치들은 센서스 추정치와 CPS 추정치들과 높은 상관관계를 보임을 알 수 있다. 또한, 합성추정치와 실직보험가입비율 추정치와는 낮은 상관성을 나타내며, 이들과 70-step 추정치들과도 상관계수의 값이 낮다. 따라서 추가적인 독립변수로써 합성추정치와 실직보험가입비율을 선택한다면 회귀식의 예측력은 증가할 것임을 알 수 있다.

<표 3.1> 변수들간의 가중 상관계수

	X_1	X_2	X_3	X_4	X_5	X_6	Y	Z	U
X_1	1.000								
X_2	0.676	1.000							
X_3	0.285	0.512	1.000						
X_4	0.682	0.961	0.574	1.000					
X_5	0.666	0.995	0.477	0.959	1.000				
X_6	0.369	0.584	0.974	0.633	0.548	1.000			
Y	0.372	0.692	0.577	0.720	0.682	0.599	1.000		
Z	0.259	0.525	0.340	0.543	0.521	0.339	0.700	1.000	
U	0.554	0.851	0.756	0.868	0.835	0.810	0.741	0.512	1.000

* Z = 1970년 5월 CPS 실업추정치

센서스 중간 해에는 회귀식의 종속변수는 CPS 1차 추출단위의 추정치가 되며, 독립변수로 센서스에서 추정한 실업률(U)을 선택할 수 있으며, 적합된 회귀식은 다음과 같다.

$$\hat{Y} = 0.010 + 0.450\,U + 0.326\,X_4 + 0.089\,X_6 \quad , \qquad (3.7)$$

Residual Mean Square = 0.835×10^{-4} ,

추정치의 표준오차 = 0.914×10^{-2} ,

$R^2 = 0.563$

만약 종속변수로 1970년 5월 CPS 실업 추정치(Z)를 사용한다면 적합된 회귀모형은 다음과 같다.

$$\hat{Z} = 0.019 + 0.422\,U + 0.430\,X_4 - 0.246\,X_6 \quad , \qquad (3.8)$$

Residual Mean Square = 2.040×10^{-4} ,

추정치의 표준오차 = 1.428×10^{-2} ,

$$R^2 = 0.291$$

모형 적합 시 센서스 시기뿐만 아니라 센서스 중간 해의 소지역의 추정치에 대한 자료가 필요하다. 센서스 중간 해에 종속변수로 CPS 추정치를 이용하고, 독립변수로 센서스 추정치를 이용한 회귀모형 사용이 가능하다. 식 (3.7)은 독립변수로 센서스 추정치를 이용한 결과를 보여준다. 회귀모형은 약 56%의 설명력을 가지며, 또한 Y를 종속변수로 갖는 다른 모형과 비교했을 때 결코 떨어지지 않는 설명력을 갖는다.

1970년 5월 CPS 실업 추정치를 종속변수로 취한 회귀식 (3.8)은 약 29%의 설명력을 갖는다. 이러한 낮은 설명력은 종속변수의 큰 분산(변동)에 기인한다.

노동력의 규모가 서로 상이한 지역들에 대한 잔차분석 결과가 <표3.2>에 주어졌다. 독립변수가 각각 (X_1 , X_3 , X_4), (X_1 , X_5 , X_6)이고, 종속변수로써 각각 1970년 센서스 추정치, 5개월간의 월별 CPS 추정치, 1970년 5월의 CPS 추정치를 이용한 회귀모형에 대한 잔차의 평균과 잔차의 표준오차이다.

<표 3.2> 잔차의 평균과 표준오차 (단위 : %점)

1970 Census Labor Force	Number of SMSA's	Census unemployment		CPS-five month average unemployment		CPS-April 1970 unemployment	
		Mean residual	S.E of residual	Mean residual	S.E of residual	Mean residual	S.E of residual
(X_1, X_3, X_4)							
1,000,000 이상	9	-0.18	0.31	-0.01	0.39	-0.00	0.54
500,000-999,999	17	-0.02	0.54	-0.06	0.66	0.24	0.88
250,000-499,999	21	0.15	0.36	0.15	1.17	-0.21	1.23
100,000-249,999	51	0.14	0.66	0.03	1.55	0.10	2.77
100,000 미만	24	0.37	0.78	0.41	1.83	-0.46	2.83
(X_1, X_5, X_6)							
1,000,000 이상	9	-0.07	0.34	0.07	0.38	0.06	0.55
500,000-999,999	17	-0.10	0.49	-0.16	0.75	0.10	0.90
250,000-499,999	21	0.06	0.39	0.06	1.23	-0.28	1.37
100,000-249,999	51	0.10	0.61	0.03	1.57	0.14	2.80
100,000 미만	24	0.11	0.71	0.19	1.75	-0.63	2.86

노동력의 규모가 작은 지역들보다는 노동력의 규모가 큰 지역들에서 추정치의 효율이 좋게 나타난다. 1970년 Cessus 추정치를 이용한 회귀모형을 제외하면, 잔차의 표준오차는 노동력의 규모가 작을수록 증가하는 경향을 보인다. Census 추정치를 종속변수로 취한 회귀모형에서 노동력의 규모가 100,000이하인 지역에서의 잔차의 표준오차는 노동력의 규모가 백만 이상인 지역보다 2배 이상 큰 표준오차를 갖는다. 5개월 간의 월별 CPS 평균에 대한 추정치를 이용한 회귀모형에서의 잔차의 표준오차를 검토해 볼 때, 1970년 5월 CPS 추정치를 이용한 회귀모형 보다는 잔차의 표준오차가 훨씬 작음을 알 수 있다.

1970년 센서스 실업률과 1970년 센서스 실업 추정치를 이용한 회귀모형 추정치 간의 차이에 대한 분포가 <표3.3>에 주어졌다. 독립변수는 (X_1 , X_5 , X_6)를 이용하였다. <표3.1>을 살펴보면 센서스 실업 추정치를 이용했을 경우 CPS 추정치를 이용한 경우보다 차이의 분포가 훨씬 대칭적임을 알 수 있다.

−1.0~1.0 %점의 구간을 살펴보면, 센서스 회귀모형의 경우는 91.9%, CPS-5-month의 경우는 75.4%, CPS-April의 경우는 73.0%가 위치함을 발견할 수 있으며, −0.50%점 아래 구간에서는 CPS 회귀식이 큰 비율로 분포해 있음을 알 수 있다(CPS-5-month의 경우 62.2%, CPS-April의 경우 59.8%가 위치함).

<표 3.3> 센서스실업률과 회귀추정치간의 차이에 대한 분포

(122개의 SMSA 이용)

Difference classes (% point)	Census unemployment		CPS-five-month average unemployment		CPS-April 1970 unemployment	
	No. of SMSA's	Percent of SMSA's	No. of SMSA's	Percent of SMSA's	No. of SMSA's	Percent of SMSA's
3.00 이상	0	0	0	0	0	0
2.00~3.00	1	0.8	1	0.8	1	0.8
1.50~2.00	0	0	0	0	0	0
1.00~1.50	6	4.9	0	0	2	1.6
0.50~1.00	14	11.5	4	3.3	5	4.1
0.25~0.50	15	12.3	5	4.1	8	6.6
0.10~0.25	19	15.6	1	0.8	3	2.5
-0.10~0.10	23	18.9	9	7.4	6	4.9
-0.25~-0.10	10	8.2	11	9.0	8	6.6
-0.50~-0.25	15	12.3	15	12.3	16	13.1
-1.00~-0.50	16	13.1	47	38.5	43	35.2
-1.50~-1.00	3	2.5	25	20.5	23	18.9
-2.00~-1.50	0	0	2	1.6	5	4.1
-3.00~-2.00	0	0	2	1.6	2	1.6
-3.00 이하	0	0	0	0	0	0

5개월 간의 월별 CPS 추정치 평균을 종속변수로 취한 회귀모형에 대한 MSE의 추정치 $MSE = E\left\{\frac{(Y_0 - \hat{Y})'(Y_0 - \hat{Y})}{n}\right\} - \frac{(n-2p-2)\,\sigma_w^2}{n}$ 을 이용하여 계산된 결과가 다음에 주어졌다. 단, Y_0 = '관측값 Y', n = '관측값의 개수', p = '독립변수의 수', σ_w^2 = 'PSU 내의 오차'를 나타낸다.

Independent Variables	*MSE*
X_1 X_2 X_3	0.405×10^{-4}
X_1 X_3 X_4	0.369×10^{-4}
X_1 X_5 X_6	0.419×10^{-4}

그러나 이러한 MSE의 수치들은 단순한 대략적인 추정결과이다. 왜냐하면 독립변수들 간의 상관계수가 상당히 높은 값을 갖고 있기 때문이다.

상기한 자료들은 다음과 같은 내용을 시사한다. 미 노동부에서 연말 공표하는 실업 추정치는 CPS의 1차 추출단위 추정치를 독립변수로 취한 회귀모형을 이용하여 개선될 수 있음을 시사한다. 추가적인 독립변수들로 70-step 추정치 외에 실직보험 가입자료와 센서스와 CPS 자료를 이용한 합성추정치 및 센서스 중간년도의 표본 자료의 추정치 등을 이용할 수 있다.

회귀모형에 의한 소지역 실업 통계의 추정은 적절한 독립변수와 종속변수의 선택이 중요한 과제이지만, 한편, 센서스와 CPS자료를 이용한 실질적인 분석 및 실업에 관련된 사회경제적 요인파악도 중요한 과제이다.

3.1.3 시계열-회귀모형

시계열-회귀모형은 현재 미국에서 주 단위 등의 소지역 실업통계를 작성하는데 적용하고 있는 방법으로 모집단의 값을 확률과정으로 생각할 뿐만 아니라 직접 표본조사 추정량을 개선하기 위해서 시계열 분석의 signal 적출 기법을 적용한 것이다.

1989년 1월 노동통계국(BLS)에서 미국의 39개 주와 콜롬비아의 1개 구역 총 40개 주에 대해 월별 고용과 실업 추정을 위한 새로운 방법을 소개하였다. CPS의 월별 표본자료에 시계열 모형을 적합시키는 방법이다.

모집단의 특성을 추정하기 위한 직접적인 방법은 표본 설계에 근거하여 대

규모 표본조사를 시행하는 것이다. 추정치들은 대 지역에 대해서는 신뢰할 만 하나 소지역에 대해서는 그러하지 못한 실정이다. 주기적인 조사의 경우 소지역에 대해서 시계열 기법이 관심을 받고 있으며, CPS 추정치는 이러한 기법을 적용시키기에 적합한 자료이다. 매달 59,000가구가 조사되어 모집단의 노동력 상태를 추정하는데, 전체 추정치 또는 인구수가 많은 11개 주에 대한 월별 추정치는 비교적 신뢰할 만 하다. 인구수가 적은 나머지 40개 주에 대해서는 월별 추정치를 그대로 사용하는 것은 바람직하지 못하다. 1989년 이전에는 40개 주에 대한 노동력 추정치는 BLS 핸드북에서 제시한 방법에 의해서 수행되었다.

새로운 추정방법은 CPS 표본 자료를 확률적으로 변화하는 노동력 시계열인 signal 성분과 noise 성분의 합으로 표현되는 모형 추정에 근거한다. 표본 설계 정보에 따라 월별 CPS 노동력 추정치는 시계열 모형에서 실업보험 자료(UI :Unemployment Insurance Data)와 경상고용통계자료(CES: Current Employment Statistics Data)를 결합하여 계산한다. 즉, 이전에 추정한 방법과는 달리 좀더 체계적인 방법으로써 보조자료와 과거와 현재의 표본자료를 모두 이용하여 표본크기가 작은데에서 발생하는 CPS 추정치의 큰 분산값을 줄이고자 하는 것이 기본적인 생각이다.

비 관측된 모집단 값의 동적인 변화와 표본오차의 자기공분산을 나타내는 모형이 주어진다면, Kalman Filter는 참값을 추정하기 위해 사용될 수 있으며, 다음과 같은 유용성을 갖고 있다. 첫째, KF는 signal성분과 noise성분으로 표현된 모형들에 대한 다양한 접근방법들을 허용한다. 둘째, KF의 반복적인 방법은 소지역의 월별 노동력 추정치를 산출하는데에 있어서 매우 효율이 좋은 알고리즘을 제공한다. 셋째, KF는 동적인 모형들에 있어서 미지인 모수들을 추정하기 위한 매우 유용한 도구이다.

(1) CPS 자료를 모형화하기 위한 시계열 방법

CPS 노동력 추정치 $y(t)$는 signal $\theta(t)$와 noise $e(t)$의 합 $y(t) = \theta(t) + e(t)$으로 나타낸다. signal 성분(노동력의 참값)은 시간에 따라 변화하는 평균 $\mu_X(t)$와 오차항 $u(t)$로 표현할 수 있다.

$$\theta(t) = \mu_X(t) + u(t) \tag{3.9}$$

$$= X(t)\beta(t) + u(t)$$

단, $X(t) = 1\times k$ 벡터(관측된 값), $\beta(t) = k\times 1$ 벡터(확률계수 벡터) 이다. 회귀계수 $\beta(t)$는 1차 자기상관회귀 과정에 따라 확률적으로 변화하는 형태를 수식으로 표현하면 아래와 같다.

$$\beta(t) = T_\beta \beta(t-1) + v_\beta(t) \tag{3.10}$$

여기에서 $T_\beta = k\times k$ 행렬(고정모수들), $v_\beta(t) = k\times 1$ 벡터(백색잡음오차)이다.

식(3.9)에서 오차 $u(t)$는 자기상관성을 고려하여 $u(t) \sim ARMA(p_u, q_u)$를 가정하여 다음과 같이 표현할 수 있다.

$$u(t) = \phi_u(L)\ \Psi_u(L)^{-1} v_u(t) \tag{3.11}$$

여기에서 $u(t) =$ 방정식의 오차, $v_u(t) = u(t)$에 대한 백색잡음오차,

$$\Psi_u(L) = 1 - \sum_{i=1}^{p_u} \Psi_u(i)\, L^i\ , \quad (\ u(t)\text{에 대한 자기회귀연산자})$$

$$\phi_u(L) = 1 + \sum_{i=1}^{q_u} \phi_u(i)\, L^i\ , \quad (\ u(t)\text{에 대한 } MA \text{ 연산자})$$

$L =$ 시차 연산자(lag operator)로써 $L^i y(t) = y(t-i)$를 만족하고, 랜덤 오차 $v_\beta(t)$와 $v_u(t)$는 평균이 0이고 서로 독립을 가정한다. 즉,

$$\begin{bmatrix} v_\beta(t) \\ v_u(t) \end{bmatrix} \sim \mathrm{ID}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} Q & 0 \\ 0 & \sigma_{v_u v_u} \end{bmatrix} \right) \tag{3.12}$$

단, $Q = Cov(v_3(t)) = Diag(\sigma_{\beta_1\beta_1}, \cdots, \sigma_{\beta_k\beta_k})$

noise 성분은 전체 모집단에서 일부를 표본 추출하는 과정에서 발생하는 오차인 표본추출오차로 취급된다. CPS는 모집단으로부터 추출된 복잡한 다단계 표본이다. 첫 단계에서 1차 추출단위(PSU's)들의 층화된 표본이 뽑히고, 다음으로 PSU 내에서 가구단위들이 층화된 집락 표본으로부터 뽑히며, 이때 매달 가구조사에서 일부 가구들은 대체 가구로 바뀌어 조사된다. CPS 추정과정은 비면접 보정인 2단계의 비율 보정과 중복표본을 고려한 합성절차로 이루어져 있다(Bureau of the Census(1978)).

$y(t)$가 노동력 특성을 나타내는 $\theta(t)$에 대한 CPS 추정치로 주어지면, 표본오차 $e(t) = y(t) - \theta(t)$의 분산과 공분산 함수는 다음과 같다.

$$\sigma_{e(t)e(t)} = D_y \, S_y^2 \,, \tag{3.13}$$

$$\gamma_{ts} = Cov(e(t) \,,\, e(s)) \,,$$

여기에서 D_y = 단순임의추출표본(SRS) 추정치의 분산에 대한 CPS 추정치 분산의 비(Design Effect),

$$S_y^2 = \frac{N(t)}{n(t)} \, \theta(t)\,(1 - P(t))\,,$$

$N(t)$ = 모집단 크기,

$n(t)$ = 표본 크기,

$$P(t) = \frac{\theta(t)}{N(t)}$$

위의 식에서 설명한 것과 같이 CPS 표본오차는 이분산적 구조와 자기상관적 구조를 동시에 갖는다. 식(3.13)은 다음과 같은 이분산성의 세가지 주요한 원인을 나타낸 식이다. 즉, 표본 재설계에서 설계효과 D_y , 표본 구간 $\frac{N(t)}{n(t)}$, 참값 $\theta(t)$와 $P(t)$의 변화가 반영되었다.

부연 설명하면, CPS는 10년 간의 센서스 자료를 사용하여 매 십년 마다 표

본 추출 구조 및 추정 절차를 갱신하기 위하여 재 설계된다. 1984-1985년 동안 비율보정과 합성과 같은 비 면접에 의한 개선된 추정 절차가 주 단위의 설계로 시행되었다. 표본조사의 재 설계보다는 오히려 주 단위의 표본 크기의 조정이 자주 있었고, 이것은 주 수준의 분산에 중요한 영향을 끼치게 되었다. 일종의 고정된 설계와 고정된 표본 크기일지라도 오차 분산은 참 노동력 크기의 함수이므로 변하게 된다. 노동력은 아주 순환적이고 동시에 계절적이므로 분산도 이와 비슷한 형태를 가진다는 것을 예측할 수 있다.

$e(t)$의 자기공분산 구조는 다음의 3가지 사실에 기인한다. 첫째, 월별 표본은 8개의 독립인 연동교체그룹들인 부 표본들로 구성되어 있다. 연동교체그룹들은 4개월 간 연속 조사되고, 8개월 동안 조사되지 않다가 다시 4개월 간 연속 조사 후 표본에서 영구히 제거된다. 이 과정에서 동일 주택단위들이 나타나므로 당연히 상관성이 존재한다. 4-8-4 연동교체 체계는 월별 추정값의 신뢰성을 높이기 위해 15개월 주기 동안 첫 달은 전월에 조사된 표본의 75%를 중복하여 월 표본으로 사용하고, 나머지는 교체표본 추출방식에 의해 새로운 표본을 조사한다. 2월과 12월을 살펴보면 중복율을 50%이다. 둘째, 연동교체 체계의 사용은 표본들에 대한 주기적 선택을 요구한다. 한 집락의 주택단위들이 연동교체그룹으로부터 영구히 제외될 때, 근처에 있는 단위들로 대체된다. 따라서 새로운 단위들도 대체될 단위들과 비슷한 특성을 가질 것이므로 같은 연동교체그룹에서 다른 가구들과도 높은 상관성을 갖게된다. 셋째, 표본 오차의 변화성은 복합추정량에 의해 영향을 받게된다. CPS 추정량의 이분산적이며 자기상관적 구조는 $e(t)$를 $e(t) = \gamma(t)\, e^*(t)$와 같은 승법적 구조로 모형화 함으로써 설명될 수 있다(Bell and Hillmer(1989)). 여기에서 $e^*(t)$는 $ARMA$과정을 따르며 상수인 분산을 갖는다.

$$e^*(t) = \phi_e(L)\ \Psi_e(L)^{-1}\ v_e(t) \qquad (3.14)$$

여기에서 $v_e(t) \sim \text{NID}(0, \sigma_{v_e v_e})$, $\sigma_{e^*(t)e^*(t)} = \sigma_{v_e v_e} \sum_{k=0}^{\infty} g_k^2$, 가중치 $\{g_k\}$는 생성함수 $g(L) = \phi_e(L)\Psi_e(L)^{-1}$로 계산된다. $e(t)$의 이분산성 성분 $\gamma(t) = \sqrt{\frac{\sigma_{e(t)e(t)}}{\sigma_{e^*(t)e^*(t)}}}$는 분산비의 제곱근이다. $e(t)$의 자기상관구조는 표본 설계에 의해 영향을 받게된다. 예를 들면, 표본 재 설계에서 복합추정량에서의 가중치들이 모두 교체되는 것과 같은 경우이다.

(2) 상태공간형식(State Space Form)과 KF 알고리즘

노동력 모형의 signal 성분과 noise 성분을 상태공간 형식에 삽입한다. 먼저 추정에는 적절치 못할지라도 융통성을 보여 주기에는 유용할 것 같은 매우 일반적인 형식을 설명하고 실질적인 적용에 있어서는 부과되어야 할 것 같은 일종의 제한들에 대해서 토의한다.

상태공간 작성에 있어서 비 관측된 signal과 noise가 상태 변수들이고, 이들의 시간에 따른 전개는 변이함수들의 집합에 의해 설명된다. 관측점 함수는 상태 변수들을 관측 표본 시계열들로 변환시킨다. 상태공간 체계에서 변이함수들은 1계의 VAR 형태를 취한다.

우리의 문제에서 관측할 수 없는 변수는 $\beta(t)$, $u(t)$, $e^*(t)$이다. 계수벡터인 $\beta(t)$는 (3.10)식과 같이 이미 적절한 형식이 있다. $u(t)$와 $e^*(t)$는 (3.11)식과 (3.14)식에서와 같이 *ARMA*과정으로 명시되나, 각각 1계 VAR형태인 $S_u(t)$와 $S_e(t)$벡터들로 변환된다. 임의의 *ARMA*(p, q)과정은 일종의 $r \times 1$ 1계 VAR형태로 변환될 수 있다는 것이 기본적인 규칙이며, 여기에서 $r = \max(p, q+1)$이다(Harvey(1981)). 변이함수들은 다음과 같이 주어지며, 여기에서 $S(t)$는 $\beta(t)$와 $S_u(t)$, $S_e(t)$로 구성되는 상태벡터이다.

$$S(t) = T(t)S(t-1) + \Gamma(t)v(t) \ ,$$

$(m\times 1)(m\times m)\ (m\times l)(l\times 1)$

$$\begin{bmatrix}\beta(t)\\ S_u(t)\\ S_e(t)\end{bmatrix}=\begin{bmatrix}T_\beta & 0 & 0\\ 0 & T_u & 0\\ 0 & 0 & T_e(t)\end{bmatrix}\begin{bmatrix}\beta(t-1)\\ S_u(t-1)\\ S_e(t-1)\end{bmatrix}+\begin{bmatrix}I_k & 0 & 0\\ 0 & \Gamma_u & 0\\ 0 & 0 & \Gamma_e(t)\end{bmatrix}\begin{bmatrix}v_\beta(t)\\ v_u(t)\\ v_e(t)\end{bmatrix},$$

$$E(v(t)\,v(t)^t)=block\ \ diagonal(Q,\ \sigma_{v_u v_u},\ \sigma_{v_{a(t)}v_{a(t)}})$$

여기에서 $$T_u=\begin{bmatrix}\Psi_u(1) & : & I_{r_u-1}\\ \Psi_u(2) & : & \\ \cdot & : & \\ \cdot & : & \\ \cdot & : & \\ \cdots\cdots & \cdots\cdot & \cdots\cdots\\ \Psi_u(r_u) & : & 0\end{bmatrix},$$

$$T_e(t)=\begin{bmatrix}\Psi_e(1) & : & I_{r_e-1}\\ \Psi_e(2) & : & \\ \cdot & : & \\ \cdot & : & \\ \cdot & : & \\ \cdots\cdots & \cdots\cdot & \cdots\cdots\\ \Psi_e(r_e) & : & 0\end{bmatrix},$$

$$\Gamma_u=\begin{bmatrix}1\\ \phi_u(1)\\ \phi_u(2)\\ \cdot\\ \cdot\\ \cdot\\ \phi_u(r_u-1)\end{bmatrix},\quad \Gamma_e(t)=\begin{bmatrix}1\\ \phi_e(1)\\ \phi_e(2)\\ \cdot\\ \cdot\\ \cdot\\ \phi_e(r_e-1)\end{bmatrix},$$

$m=k+r_u+r_e$,

$k=$ 회귀변수의 수,

$l=k+2$,

$r_u=\max(p_u,\ q_u+1)$, p_u와 q_u는 $u(t)$의 *ARMA* 모수,

$r_e=\max(p_e,\ q_e+1)$, p_e와 q_e는 $e(t)$의 *ARMA* 모수.

관측점 함수는 벡터 $H(t)$를 선택하여, signal성분과 noise성분을 만들기 위

한 상태변수들의 일차결합을 취한다.

$$Y(t) = H(t)S(t) = \theta(t) + e(t) \ , \qquad (3.15)$$

여기에서 $H(t) = [\, X(t) | 1 | 0_{r_u - 1} | \gamma(t) | 0_{m-k-r_u-2} \,]$,

$$\theta(t) = H_\theta(t)\, S(t),\ e(t) = H_e(t)\, S(t),$$

$$H_\theta(t) = [\, X(t) | 1 | 0_{m-k-1} \,],\ H_e(t) = [\, 0_{k+r_u} | \gamma(t) | 0_{r_u-1} \,].$$

비 관측된 signal과 noise성분들의 상태공간형태가 주어진다면, KF는 signal과 noise를 추정하기 위한 방법을 제공한다. 이러한 알고리즘을 설명하기 위해 시간 $t-j$까지 관측된 자료에 대한 $S(t)$의 조건부 평균 및 공분산 행렬을 다음과 같이 표현하자.

$$S(t|t-j) = E(S(t) \,|\, Y_{t-j}, \cdots, Y_1),$$

$$P(t|t-j) = E\{ (S(t) - S(t|t-j))(S(t) - S(t|t-j))^t \,|\, Y_{t-j}, \cdots, Y_1) \}$$

또한 바로 전까지의 값들이 주어진 경우의 표본 추정치 $Y(t)$의 예측값을 $Y(t|t-1) = H(t)\, S(t|t-1)$, $Y(t)$ 의 분산을 아래와 같이 표현하자.

$$E(Y(t) - Y(t|t-1))^2 = H(t)\, P(t|t-1)\, H(t)^t = f(t|t-1).$$

t번째 관측점까지의 (t번째 관측점은 제외) 자료에 근거한 $S(t)$의 추정치가 주어진다면, 최근의 자료를 고려한 $S(t)$의 추정량은 $S(t|t-1)$과 최근의 표본 추정치 $Y(t)$의 가중평균으로 다음과 같이 표현되며,

$$S(t|t) = (I - K(t)\, H(t))\, S(t|t-1) + K(t)\, Y(t)$$

$$= S(t|t-1) + K(t)\, (Y(t) - Y(t|t-1))$$

공분산 행렬은

$P(t|t) = (I - K(t)\, H(t)^t)\, P(t|t-1)$로써, $S(t|t-1) = TS(t-1|t-1)$와 $P(t|t-1) = TP(t-1|t-1)\, T^t + \Gamma E(v(t)v(t)^t)\Gamma^t$ 로부터 반복적으로 추정된

다. 여기에서 가중벡터 $K(t)$(gain of filter)는 $K(t) = \frac{P(t|t-1)H(t)^t}{f(t|t-1)}$ 로 표현되며, $K(t)$의 원소들은 $P(t|t)$의 대각원소들의 합을 최소화하여 결정한다(Gelb(1974)).

KF 방정식을 이용하여, 시간 t에서 관측된 표본 추정치는 signal성분과 noise성분으로 분해된다.

$$Y(t) = \theta(t|t) + e(t|t) ,$$

여기에서 $\theta(t|t) = \theta(t|t-1) + h_\theta(t)\ \tilde{Y}(t)$,

$$e(t|t) = e(t|t-1) + (1-h_\theta(t))\ \tilde{Y}(t),$$

$$\tilde{Y}(t) = Y(t) - Y(t|t-1),$$

$$h_\theta(t) = H_\theta(t)\ K(t)$$

$$= \frac{\left\{ Var(\frac{\theta(t)}{\theta(t-1)}) + H_\theta(t)\ TP(t-1|t-1)\ T^t H(t) \right\}}{f(t|t-1)},$$

$$1-h_\theta(t) = H_e(t)\ K(t) \ = \frac{\{\sigma_{e(t)e(t)} + H_e(t)\ TP(t-1|t-1)\ T^t H(t)\}}{f(t|t-1)} ,$$

$$Var(\frac{\theta(t)}{\theta(t|t-1)}) = \sum_{i=1}^{k} X_i(t)^2 \sigma_{\beta_i\beta_i} + \sigma_{v_\theta v_\theta} ,$$

$$\sigma_{e(t)e(t)} = \gamma(t)^2 \sigma_{v_e v_e} .$$

가중치 $h_\theta(t)$는 예측오차 $\tilde{Y}(t)$를 signal과 noise성분으로 분해하며, 이것은 KF가 signal성분의 최소평균제곱오차 추정치를 만들기 위하여 시계열 추정량 $\theta(t|t-1)$와 최근의 표본 추정치 $Y(t)$를 결합하는 방법을 설명한다. $\theta(t|t-1)$의 $Y(t)$에 대해서 보정되는 양은 이분산성의 표본오차 분산 $\sigma_{e(t)e(t)}$와 상대적으로 비교하여 시계열 분산성분 $Var(\frac{\theta(t)}{\theta(t-1)})$의 크기의 함수로 표현된다. $\sigma_{e(t)e(t)}$의 값이 크면 $h_\theta(t)$의 값이 작게되며, 따라서

$\theta(t|t)$를 이끌어 내는데 있어서 시계열 예측치 $\theta(t|t-1)$에 대한 작은 보정만이 일어난다. 역으로 만약 표본 분산이 작다면 $\theta(t|t)$는 최근의 표본 추정치 $Y(t)$와는 아주 다른 값이 될 것이다.

KF는 반복적인 방법으로 상태벡터 $S(t)$의 최소평균제곱오차를 제공하며, 또한 KF는 새로운 자료가 각 주기에서 이용될 수 있는 실시간 상황에 적합하다. 그러나 자료가 시간 t 이후에 이용될 수 있을 경우에는 추정치 $S(t|t)$는 이러한 새로운 정보를 반영시키지는 못할 것이다. 왜냐하면 KF는 시간 t의 앞쪽으로만 이동하기 때문이다. 이 전 주기의 추정치들에 대한 부분적인 최적화는 평활을 통해 쉽게 조정될 수 있다.

평활의 방법은 두 가지 형태의 KF 추정량을 결합하는 방법이다. 첫 번째 형태의 KF 추정량은 이 전에 설명했던 것과 같이 전방 Filter 추정치로써 시간 t에서 모든 과거 표본자료와 현재 표본자료에 의해 추정되는 $S(t|t)$이다. 두 번째 형태의 KF 추정량은 후방 Filter 추정치로써, 표본주기의 끝에서 출발하여(가령 $t=n$에서 출발) 처음까지 진행하며, 미래의 자료에만 근거하여, 각 시간 t에서의 예측치들을 이끌어 내는데 이러한 추정량을 $S(t|t+1)$로 나타낸다. 이 때 최적인 평활된 추정량은 두 추정량의 평균제곱오차 값들의 비율로 결합되어 만들어지는데 다음과 같다.

$$S(t|n) = P(t|n)\left\{ \frac{S(t|t)}{P(t|t)} + \frac{S(t|t+1)}{P(t|t+1)} \right\},$$

$$\text{단, } P(t|n) = 1/\left\{ \frac{1}{P(t|t)} + \frac{1}{P(t|t+1)} \right\}$$

한편, $S(t|n)$의 공분산 표현식으로부터

$\frac{1}{P(t|n)} = \frac{1}{P(t|t)} + \frac{1}{P(t|t+1)}$을 얻을 수 있고,

$P(t|n) - P(t|t)$는 음 반정치(negative semidefinite)임을 알 수 있다. 따라서 $S(t)$의 평활된 추정량은 전방 Filter추정량보다는 훨씬 좋은 추정량이

되며, 이러한 이유 때문에 노동력 추정치들은 평활 알고리즘을 이용하여 만들어진다.

3.1.4 보충 설명

상태공간 형식은 signal성분을 세분화하는데 상당한 유연성을 허용하며, 특별한 경우들로써 표본조사 방법에 근거한 다음의 2가지 형태의 모형을 포함한다.

만약 $Q = Cov(v_\beta(t)) = 0$ 이고 $e(t)$와 $u(t)$가 백색잡음오차 변수이면, 시스템은 Ericksen(1974)의 표본회귀모형이 된다. 이 경우 signal 적출 문제는 관측된 표본자료에 대해 가중 최소제곱방정식을 적합시켜 해결된다.

$\beta(t) = 0$, $Q = 0$일 경우 Wiener-Kolmogorov의 signal 적출 이론에 근거한 모형을 얻는다. 회귀식의 평균은 없어지고 signal은 공분산 정상과정이 된다. 추가적으로 $e(t)$과정의 분산과 $ARMA$ 모수들이 상수인 경우에는 $e(t)$가 공분산 정상과정이 된다.

Scott와 Smith(1974)는 공분산이 추정되어야 하는 조사자료에 대해 전통적인 signal 적출 방법을 채택하였다. Bell과 Hillmer(1987a)는 signal 과정에서 비 정상성을 다루는 방법들을 토의하였다.

앞서 소개되었던 내용에서는 signal의 비 정상성 문제를 회귀변수들과 회귀계수들의 확률적인 변화로 다루었다. 회귀계수들의 움직임을 통제하는 (7.10)식의 추이 방정식은 다양한 형태를 수용할 수 있다(Los(1985)). 이러한 회귀계수들은 독립인 랜덤워크를 따르는 것으로 명시한다(즉, $T_\beta = I$ 이고 Q가 대각행렬임을 명시).

CPS 자료를 이용한 연구는 많은 연구가 진행되고 있지는 않다. Hausman과 Watson(1985)은 CPS 4-8-4 연동표본교체와 합성절차를 통합시킨 전 지역의 십대의 실업률 시계열에 대한 오차과정의 $ARMA(1, 15)$모형

을 개발했다. Bell과 Hillmer(1987b)는 Train(1978)등에 의해 추정된 design에 근거한 자기공분산의 근사로써 $ARMA(1,1)$모형을 개발했다.

CPS와 같은 복잡한 조사에 대해서 공분산 추정치를 산출하는 것은 대규모의 자료가 사용되어 비용이 많이 든다. 최근에는 필요로 하는 모든 자료를 이용하지 않고 이러한 추정치를 산출하는 방법이 모색되고 있다.

설계에 근거한 표본오차 공분산 계산의 어려움 때문에 방정식에서의 오차들의 효과와 표본오차의 효과를 추정하지 않고 회귀방정식을 적합시키기도 한다. 만약 두 성분 오차들이 $ARMA$과정이면, 이 때 합도 $ARMA$과정이 된다(Granger and Morris(1975)). 즉,

$$u(t) \sim ARMA(p_u, q_u), \quad e(t) \sim ARMA(p_e, q_e) \text{ 이면}$$

$$w(t) = u(t) + e(t) \sim ARMA(p, q)$$

여기에서 $p \leq p_u + p_e$, $q \leq \max(p_u + q_e, p_u + q_e)$. KF는 회귀성분과 총 오차를 적출하기 위해 사용될 수 있다.

3.1.5 실업률 추정에 적용

CPS외의 자료로써 미 연방-주의 UI 체계로부터 생산된 실직보험가입자료(UID)와 비농인구에 대한 경상고용통계조사(CES)자료가 있다. 이러한 자료들은 1960년대 초 이래로 주의 특정 추정치를 생산하기 위하여 핸드북 방법에서 이용되었던 자료들이다.

UI자료와 CES자료로 설명되지 않는 순환적이며 계절적인 노동력의 움직임을 통제하기 위하여 표본오차의 영향을 줄이는 방법으로 선택된 CPS자료로부터 변수들이 구성된다. 이러한 방법은 구성된 변수들의 오차 문제를 다루기 때문에, 즉 계수들보다는 오히려 종속변수의 참값을 추정하는 데에 초점이 맞추어져 있기 때문에 기존의 방법과는 다르다.

설명변수들로 주의 특유한 CPS자료를 이용하는 것이 바람직하나 그렇게

하기 위해서는 변수들이 갖고 있는 오차를 명확하게 설명할 수 있는 모형이 필요하게 된다. 보험가입 실직자 수에 대한 월별 주단위 자료가 CPS와는 독립적으로 얻을 수 있는 실직자에 대한 최신정보이다. 이러한 자료가 실업률 모형 개발을 위한 출발점이 된다. 실직보험자료는 UI혜택을 신청하고 있는 노동자들의 수를 완전히 집계한다. 일반적으로, 해고되어 주의 특정한 재정상의 적격 기준을 충족시킨 노동자들에게만 혜택이 돌아간다. 대조적으로 CPS에서 이용되는 실직의 개념은 조사기간 동안 직업을 갖고 있지 않은 모든 사람들, 해고되어 직업을 찾고 있는 사람 또는 일시 해고되어 대기발령중인 사람들을 모두 포함한다. 미국의 경우 실직보험자료에서 실직자에 포함되지 않은 실직그룹들을 살펴보면 <표3.4>와 같다.

<표 3.4> 실직보험자료(UID)에서 집계 안된 비고용 그룹

1. 다음의 범주에 해당되는 실직자
   a. Exhaustees : 수혜 권리를 다 써버린 노동자들
   b. 재정적 부적격자 : 주의 적격 요구기준을 충족시키지 못하는 우선 고용자 또는 소득을 갖고 있는 노동자들
   c. 유예된 비 신청자 : 실직시기의 처음에 혜택을 신청하지 않은 적격한 노동자들
2. 노동력 신규 유입자 : 최근의 실직기간 전에 노동인구에 편입되지 않았던 노동자들
3. 이직자 : 이 전의 직업을 그만두고 다른 직업을 찾고 있는 노동자

UI가 집계 않는 부분들 중에서 신규 유입자가 차지하는 비율이 가장 크며, 다니는 직장을 그만두고 다른 직업을 찾는 이직 희망자들은 적어도 그 기간 만큼은 보험상의 혜택을 받지는 못한다. 만약 UI혜택을 받지 못하는 실직자의 상대적인 규모가 시간에 따라 안정적이라면, 보험 지불요구율은 전체 실업률 추정에 대용이 될 수 있다. 실제로 노동시장에서 특히 실직자와 신

규 유입자 간의 실업률 분포는 순환적이며 계절적인 특성변화를 나타낸다.
먼저 실업률 분포에 있어서 계절적 변화를 고려해 보자. 가장 중요한 현상은 실직자와 유입자는 매우 다른 계절적 형태를 띤다는 점이다. 청년과 여성이 유입자의 대부분을 차지하고, 청년층의 실업은 휘발성의 계절적 형태를 띠는데, 이는 학기가 끝나면서 신규 노동력이 유입되고 학기 시작과 함께 빠져나가는 청년층의 노동력의 변화가 반영되었기 때문이다. 이와는 대조적으로 성인 남성의 실직에 있어서 가장 일반적인 원인은 자동차 산업과 건설산업과 같이 산업의 연간 생산 순환주기에 영향을 받아 계절적인 해고와 재고용이 발생한다. <표3.5>는 CPS 신규유입 및 실직률의 계절적 형태(40개주에 대한 평균값)와 전체 비율에 대한 이러한 항들의 순수효과 간의 차이를 나타내며, 또한 실직보험 가입자에 대한 계절적인 형태를 보여준다.
여기에서 100보다 큰 값은 평균실업률보다 큰 달을 말한다. 신규 유입률은 겨울에는 평균실업률보다 낮고 여름에는 평균실업률보다 높다. 반면, 실직률은 반대의 형태를 보인다. 각 그룹에서의 큰 계절 실업률은 전체 실업률에 강한 영향을 미친다. 실직자와 신규 유입자는 경기순환과 다른 형태를 띤다는 점에서 수치적으로 중요하며, 전체 실업률의 약 반 정도를 설명한다. 이직자는 전체의 약 15%를 설명하므로 수치적으로 덜 중요하다. 늦여름과 가을에 구별되는 계절적 형태를 나타낸다. 실직보험 지불요구율은 실직자의 계절적 형태를 따르나 신규 유입 또는 이직자의 계절적 형태를 따르지는 않는다.

<표 3.5> 40개 주에 대한 실업 계절 요인들에 대한 평균(1979-85)

월	CPS 실업률				UI 지불요구율
	전체	신규유입자	실직자	이직자	
1월	110.6	99.1	120.9	104.3	131.2
2월	110.9	98.1	126.1	100.7	133.8
3월	105.3	96.2	115.7	95.5	120.8
4월	97.0	90.7	103.5	91.3	104.0
5월	93.2	96.4	91.0	92.8	91.0
6월	106.0	127.2	89.6	93.6	85.8
7월	98.7	106.0	90.8	101.0	95.9
8월	98.5	102.3	92.1	110.7	90.2
9월	95.5	102.7	84.7	112.8	77.6
10월	93.6	98.3	88.0	107.0	79.3
11월	94.9	94.5	93.5	101.3	87.7
12월	95.9	88.4	104.1	89.5	102.4

* <표3.5>에서 계절 요인들은 X-11로 계산됨. CPS 실업률의 분모는 각각의 범주에 대해서 CPS 고용과 실직인원의 합이고, UI 지불요구율에서 분모는 CES고용의 총계임.

다음의 <표3.6>은 경기순환과 관련한 실직률의 분포의 변화를 나타낸다. 경기후퇴 기간 중에는 노동력 요구는 떨어지고 실직자는 증가한다. 따라서 실직보험 지불요구율은 충분히 순환적인 지표가 된다. 그러나 실직보험 지불요구는 실직자의 순환적인 변화를 전적으로 반영하지는 못한다. 경기후퇴의 후반부 쪽에서는 실직기간이 길어져서 UI 혜택을 모두 써버린 노동자들이 증가한다. 또한 일단 재 고용되면 이러한 노동자들이 차후의 실직기간 동안 수혜에 대한 자격을 얻기 위한 고용신뢰도와 충분한 임금을 얻기까지는 얼마간의 시간이 걸린다. 중요한 사실은 실직자들에 대한 UI 범위를 살펴보면 오랜 기간에 걸쳐 계속해서 줄어든 적이 있었다는 점이다. <표3.6>에서 이러한 사실을 확인할 수 있다. 경기후퇴 기간 동안 증가하는 대신 UI

값이 줄어들었다.

Burtless et al.(1984)와 Corson et al.(1988)의 연구에 의하면 이러한 현상은 경제적인 변화나 인구 통계적인 변화와는 무관하며 UI 대상자들이 혜택을 적용 받지 못한 이유 때문으로 설명한다. 정책의 변화가 주요한 원인으로 주목된다.

이상의 토의는 대표적인 주의 움직임을 설명한 반면, 모형화 과정에서 설명되어야만 하는 중요한 주 간의 차이점들이 존재한다. UI 자료에서 보면 주의 적격요구 기준, 수혜기간, UI 적용범위에 대한 행정관례에서의 변동사항들이다. 순환적이며 계절적인 움직임에 있어서 실제적인 차이들은 회귀계수와 모형의 모수에 영향을 끼친다.

<표 3.6> 실직범주들의 상대적인 크기(40개 주 평균)

Year	U.S 실업률	Percent of Total CPS Unempoyment				
		CPS실직자 UI지불요구 (%)	지불요구	실직자	신규유입자	이직자
76	7.7	93.1	36.9	43.1	41.9	14.8
77	7.1	86.5	34.5	41.3	43.6	14.9
78	6.1	87.6	31.4	37.7	46.1	16.2
79	5.8	87.5	32.8	39.4	44.4	16.1
80	7.1	76.7	35.8	47.6	38.3	14.0
81	7.6	68.1	31.9	48.8	38.4	12.7
82	9.7	61.8	34.3	56.3	34.5	9.1
83	9.6	50.7	27.7	55.1	35.5	9.2
84	7.5	52.2	25.7	50.1	38.9	10.9
85	7.2	56.6	26.9	48.3	40.1	11.5
86	7.0	60.3	27.9	47.7	38.9	13.3
87	6.2	57.0	25.3	45.6	40.2	14.0
88	5.5	62.3	26.4	43.9	40.4	15.5

UI 적용에 있어서 변화는 <표3.7>에 주어졌다. <표3.7>은 주별 전체 CPS실업자에 대한 UI 지불요구율의 연간 평균값들의 전체평균(%), 주 내에서의 CV값, 연간 평균값들 중 최소값과 최대값을 나타냈다.

<표 3.7> 주별 전체 CPS 실업자에 대한 UI 지불요구자(1976-87)

주	UI지불 요구(%)	CV	최소값	최대값	주	UI지불 요구(%)	CV	최소값	최대값
AL	25.5	23.7	17.5	37.5	MT	32.2	21.6	23.1	44.2
AK	57.1	21.2	43.1	83.0	NE	30.0	18.7	20.7	43.9
AZ	24.3	13.8	19.8	29.5	NV	34.7	19.7	25.3	48.2
AR	28.1	21.6	20.2	38.5	NH	26.4	17.4	19.2	34.7
CO	23.6	10.9	20.2	28.1	NM	24.2	10.5	20.4	28.1
CT	37.5	17.0	29.3	48.9	ND	31.1	14.7	24.3	36.7
DE	29.9	17.1	22.8	38.0	OK	27.5	21.7	19.4	39.0
DC	26.6	10.7	21.9	30.3	OR	37.6	16.7	28.8	48.4
GA	25.8	15.0	21.2	33.4	RI	51.2	12.7	39.5	61.4
HI	32.9	9.5	29.0	39.8	SC	26.3	18.1	19.8	33.9
ID	31.6	15.4	24.7	39.5	SD	22.5	30.6	14.7	35.0
IN	24.4	17.5	19.4	33.5	TN	27.2	26.2	18.1	41.0
IA	30.3	20.9	22.6	42.1	UT	31.4	25.7	19.9	41.1
KS	35.5	13.6	29.5	45.9	VT	40.0	9.7	33.9	46.3
KY	29.2	29.7	17.2	40.0	VA	17.6	20.1	13.3	24.7
LA	27.9	20.6	20.0	37.6	WA	36.1	15.7	30.1	49.1
ME	35.8	11.2	28.7	41.1	WV	32.8	24.7	21.1	42.7
MD	28.5	12.2	23.6	34.0	WI	34.2	24.1	24.8	47.1
MN	33.7	18.7	24.0	44.4	WY	28.7	28.7	20.7	49.7
MS	26.1	19.7	20.1	35.3	ALL	31.0	23.2	17.6	57.1
MO	32.8	19.5	24.3	44.2					

비율모형의 회귀성분의 일반적인 형태는 다음과 같이 주어진다.

$$\text{실업률} = \text{Intercept}\ (t) + \beta_1(t)\ (\text{지불요구율}) \qquad (3.16)$$
$$+ \beta_2(t)\ (\text{인구 대 고용비(EPR)}) + \beta_3(t)\ (\text{신규유입률})$$

여기에서 지불요구율 = $\frac{contiued\ claims\ w/o\ earnings}{CES\ employment} \times 100$,

인구 대 고용비 = $\frac{CPS\ employment}{CPS\ 16+\ population} \times 100$,

신규유입률 = $\frac{CPS\ 고용\ 신규유입자}{CPS\ 신규유입자 + CPS\ employment} \times 100$.

지불요구율은 UI 혜택을 받고 있는 실직자들의 상대적인 규모를 나타내는 측도이다. 인구 대 고용비는 보험 지불요구 인원에 포함되지 않은 실직자들을 나타내는 측도이다. 숙련된 노동자들에 대한 상대적으로 고정된 노동력 관계비가 주어진다면, 경기순환 동안의 그들의 실직은 역으로 인구 대 고용비와 관계가 될 것이다. 이직자의 수에 영향을 미치는 노동요구에 있어서의 계절적 변동을 인구 대 고용비에서 찾아낼 수 있다. 즉 여름에 정점을 취하고 가을에 떨어지는 계절적 변동을 보인다. 이러한 주기 동안 수많은 계절적 직업들이 생기며, 이직에 기인한 실직이 증가한다.

대부분의 주에서는 고용 대 인구비의 분모의 값을 CPS 고용자료를 이용하여 계산한다. 주의 CPS 고용자료는 표본추출오차의 영향을 받을 가능성이 있지만, CV의 값은 주의 CPS 추정치 보다는 5~6배정도 작다. 소수의 주에서 인구 대 고용비의 계산에서 CES자료를 이용한다. 대부분의 주들에 대해서 CES는 CPS 고용측도와는 어느 정도 다른 계절적 형태를 갖고 있으며 실직률의 참값과는 높은 상관성을 갖고 있지는 않은 것 같다. 표본추출오차의 효과를 줄이기 위하여 CPS 신규 유입률 변수는 주보다는 훨씬 큰 지리학적 지역(전국 또는 센서스 지역)에 대해서 계산된다. 몇몇 경우에 있어서는 주의 CPS 신규 유입률의 3-month 이동평균이 사용된다.

40개 주의 각각에 대한 모형들이 1976-87년 동안의 월별 CPS 실직률의 시계열 자료에 적합되었다. 전에 설명한 것과 같이 CPS 표본자료는 다음과 같이 signal성분과 noise성분으로 표현된다.

$$Y(t) = \theta(t) + e(t) ,$$

$$\text{단, } \theta(t) = X(t)\,\beta(t) + u(t).$$

여기에서 계수들은 다음과 같은 랜덤워크를 따른다.

$$\beta(t) = \beta(t-1) + v_\beta(t)$$

$u(t)$와 $e(t)$의 효과는 분리하여 추정하지 않기 때문에 관측된 값들은 다음과 같이 표현할 수 있다.

$$Y(t) = X(t)\,\beta(t) + w(t) \quad ,$$

$$\text{단, } w(t) = u(t) + e(t) \sim ARMA(p\,,\,q)$$

$S_w(t)$를 $\max(p\,,\,q+1)$과 같은 계(order)를 갖는 $w(t)$의 상태벡터라 하면, 상태공간 모형은 다음과 같은 전이방정식(transition equation)을 갖는다.

$$S(t) = \begin{bmatrix} \beta(t) \\ S_w(t) \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & T_w \end{bmatrix} \begin{bmatrix} \beta(t-1) \\ S_w(t-1) \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & \Gamma_w \end{bmatrix} \begin{bmatrix} v_\beta(t) \\ v_w(t) \end{bmatrix},$$

$$Y(t) = [\,X(t)\;1\;0\;\cdots\;0\,]\,S(t)\ .$$

이 시스템의 모수들은 다음과 같다.

$$Cov(v_\beta(t)) = Diag(\sigma_{\beta_1\beta_1}\,,\,\cdots\,,\,\sigma_{\beta_k\beta_k})\ ,$$

$$Var(v_w(t))\ ,$$

T_w는 p AR 모수들을 갖고, Γ_w는 q MA 모수들을 갖는다.

이러한 모수들은 우도함수의 새로운 형태를 이용하여 추정된다 (Schweppe(1965)). 만약 백색잡음오차 $v_\beta(t)$와 $v_w(t)$가 정규분포를 따른다면, 1-step 앞의 예측오차들인 $\hat{Y}(t)$는 독립인 $N(0\,,\,f(t|\,t-1))$을 따르는 확률변수가 된다.

관측된 표본오차의 결합확률은 각각의 밀도함수의 곱으로 표현된다. 만약 상태벡터가 l개의 비 정상원소를 갖고 있다면, 결합밀도함수는 처음 l개의 관측점에 대한 조건부 함수로 표현되며, 미지인 모수들인 Ω의 함수로써 로그우도함수

$$L(\Omega) = -\frac{1}{2}\left\{\sum_{t=l}^{n} \log f(t|t-1) + \frac{Y(t)^2}{f(t)}\right\}$$

는 상수의 범위내에 있게된다. 만약 Ω가 주어진다면, 초기값 $S(l+1|l)$, $P(l+1|l)$을 계산하기 위한 처음의 l개의 관측값을 이용하여 $L(\Omega)$를 결정하는데, 이때 KF반복법이 사용된다. 일반적으로 이러한 방법은 어려운 비선형 최적화 문제에 해당된다. 초기에 우리는 $Cov(v_\beta(t)) = qD$ 로 나타냈다. 여기에서 q는 상수이고, D는 사전에 명시된 상수들의 대각행렬이다. 또한 $w(t)$에 대한 1계인 AR모형으로 출발하여, 두 모수들의 추정문제를 다루었다. 계수들과 AR 모수 값의 변동의 정도에 대한 개략적인 추정치들을 계산하였고, 몇몇 경우에서 이러한 추정치들은 Watson과 Engle(1983)에 의해 개발된 EM-scoring 알고리즘의 초기값들로 이용되어 좀 더 다듬어지게 되었다. 이러한 알고리즘은 계수의 변화와 관측오차를 동시에 고려하는 일반적인 자기회귀 구조에도 이용된다.

모수들에 대한 초기값들은 합리적인 반복횟수 내에서 수렴값을 얻기 위해 매우 중요하다. 계수들의 표준편차($\sqrt{\sigma_{\beta_i\beta_i}}$)는 관측오차의 표준편차의 약 0.6%정도였다. 초기값들로 이러한 사실들을 이용한다면 EM 알고리즘은 일반적으로 3~6회의 반복 후 수렴하게 된다. Harvey와 Phillips(1979), Ansley와 Kohn(1985), Bell과 Hillmer(1987a)에 의하면 KF를 초기화하는 많은 방법들이 있다.

3.2 영 국

3.2.1 개 요

정부 공식통계는 경제 정책, 재원 분배 및 정책 결정 등에 참고 자료로 이용된다. 대 영역에 대해서는 공식 통계의 정보가 이용자들에게 제공되나 소지역에 대해서는 그렇지 못한 실정이다. 최근 영국 내에서는 소지역 통계

작성에 대한 요구가 꾸준히 제기되고 있고 특히 노동시장 동향에 대한 측도 개발이 시급히 요구되고 있다. 노동력 조사(LFS)는 노동시장 정보 파악에 중요한 역할을 담당하고 있으나 직접조사 추정값들은 소지역 추정에는 한계를 안고 있다. 이 논문은 LFS로부터 소지역 추정의 신뢰도를 향상시킬 수 있는 모형 기반 추정량들을 소개한다.

영국의 LFS는 세 달을 주기로 연속 조사가 실시된다. LFS는 약 60,000 조사가구 단위를 갖는 대규모 조사로써 16세 이상의 약 150,000명의 인구에 대해 조사가 이루어진다. 영국의 LFS는 국제노동기구(ILO)의 요구조건을 만족하도록 표본설계되어 있으며 이를 통해 실업통계가 작성된다. LFS 표본설계에서 표본은 단순임의추출로 추출되며 주로 국가 수준의 추정값을 생산하도록 설계되어 있다. 일년에 한번 소지역 단위인 UA지역(Unitary Authority)과 LAD지역(Local Authority)에대한 추정값들이 작성된다. 이 연구에서 소개되는 내용은 현재 영국 통계국(ONS)에서 진행하고 있는 소지역 추정법과 밀접한 관계가 있다.

소지역 추정법은 소지역에 대한 직접 조사 추정값들이 신뢰성에 문제가 있거나계산될 수 없을 때 이용할 수 있는 통계적인 기법으로써 인근지역의 보조정보를 빌려 소지역의 특성값을 추정하는 간접 추계 방법이다. 이 연구에서 이용한 주요 보조정보는 실업보험을 청구한 사람들의 수이다. 실업보험 자료는 행정 시스템에 의해 획득되기 때문에 표본오차가 없고 지역적 범주로 또는 성별-연령대별 범주들로 다양하게 분류될 수 있다. 실업보험 지급 청구자 수와 ILO 실업자 수와는 시기에 따라 약간의 차이는 있지만 강한 상관성을 나타낸다. 시기에 따라 발생하는 차이는 주로 행정 시스템의 변경 또는 경제 사이클의 변화 등에 기인한다. ONS는 Southampton 대학과 연계하여 LFS자료와 실업보험 청구자 수의 자료를 결합하여 소지역 추정값의 신뢰성을 확보할 수 있는 연구를 진행하고 있으며, 특히 UA 또는 LAD 지역

에 대한 실업자 수를 추정하는 SPREE 방법(Purcell and Kish, 1980), 로지스틱 모형에 근거한 일종의 변형된 Fay-Herriot 방법(1979)과 Multi-level 모형화 방법(Goldstein, 1995)과 같은 세 가지 추정방법에 대해 연구를 진행하고 있다. SPREE 방법보다는 로지스틱 모형에 근거한 변형된 Fay-Herriot 방법과 Multi-level 모형화 방법이 더 좋은 효율을 나타내며, 여기에서는 변형된 Fay-Herriot 방법에 초점을 맞추어 소개한다.

3.2.2 소지역 추정 방법

앞으로 소개되는 수식에서 첨자 i와 j는 각각 UA 지역과 LAD 지역에 대한 성별-연령대별 그룹을 나타내며, 첨자 g와 h는 각각 UA 지역과 LAD 지역들을 나타낸다. 표본 자료는 LFS 추정값들과 각 지역 내에서 성별-연령대별 그룹으로 분류된 각 셀들에 대한 실업보험 청구자 수의 자료들로 이루어져 있다.

N_{ig}를 셀 (i, g)에서의 인구 총계, U_{ig}를 같은 셀에서의 실업자 총계라 할 때, 이 셀에서의 실업률은 $Z_{ig} = U_{ig}/N_{ig}$이다. 일반적으로 실업률 Z_{ig}는 g번째 지역의 특성값들에 의해 결정된다. Z_{ig}의 기대값과 분산을 각각 $E(z_{ig}) = \pi_{ig}$, $Var(z_{ig}) = \pi_{ig}(1 - \pi_{ig})/N_{ig}$라 하자. g번째 지역의 특성값들이 $E(Z_{ig})$의 값에 미치는 영향을 열거하기 위해 로지스틱 모형이 이용되었다. 이용된 로지스틱 모형은 $\text{logit}(\pi_{ig}) = \boldsymbol{x}_{ig}^T \boldsymbol{\beta}$ 이다. 여기에서 벡터 $\boldsymbol{x}_{ig}$는 소지역 g에서 i번째 성별-연령대별 그룹에 대한 속성들을 나타내며 알고있는 값이다.

N_{ig}^*를 셀 (i, g)에서의 인구 총계에 대한 LFS 추정값이라 하고, U_{ig}^*를 실업자 수에 대한 추정값이라 할 때, LFS 실업률 추정값은 $Z_{ig}^* = U_{ig}^*/N_{ig}^*$로 나타낼 수 있다. 성별-연령대별 그룹들이 합리적으로 정의되어 그룹 내에서

추출된 조사단위들에 대한 표본 가중치들에서 변동이 거의 발생하지 않는다고 가정할 수 있다면 셀 (i, g)에서의 실업률에 대한 LFS 추정값들은 표본 실업률로 근사될 수 있다. LFS 표본은 단순임의추출 표본이므로 다음 식들이 성립한다.

$$E(Z_{ig}^{*} \mid Z_{ig}) = Z_{ig} \ , \tag{3.17}$$

$$Var(Z_{ig}^{*} \mid Z_{ig}) = [(N_{ig} - n_{ig})/(N_{ig}-1)][Z_{ig}(1 - Z_{ig})/n_{ig}] \tag{3.18}$$

$$= Z_{ig}(1 - Z_{ig})/n_{ig}^{*}$$

여기에서 $n_{ig}^{*} = n_{ig}(N_{ig}^{*}-1)/(N_{ig}^{*}-n_{ig}^{*})$이고, n_{ig}는 셀 (i, g)의 LFS 표본크기를 나타낸다. 주어진 Z_{ig}에 대해 Z_{ig}^{*}와 x_{ig}의 독립성을 가정한다면 위의 식들은 다음과 같이 주어질 수 있다.

$$E(Z_{ig}^{*} \mid Z_{ig}) = E[E(Z_{ig}^{*} \mid Z_{ig}, x_{ig}) \mid x_{ig}] = E(Z_{ig} \mid x_{ig}) = \pi_{ig} \ , \tag{3.19}$$

$$Var(Z_{ig}^{*} \mid x_{ig}) = E[Z_{ig}(1-Z_{ig})/n_{ig}^{*} \mid x_{ig}] + Var(Z_{ig} \mid x_{ig}) = \pi_{ig}(1-\pi_{ig})/n_{ig}^{**} \tag{3.20}$$

여기에서 $n_{ig}^{**} = n_{ig}^{\infty}[1+(n_{ig}^{\infty}-1)/N_{ig}^{*}]^{-1}$이다. 일반적으로 n_{ig}는 N_{ig}에 비해 상대적으로 작은 값을 갖기 때문에 식 (3.19)와 (3.20)은 Z_{ig}^{*}에 대한일종의 근사 이항 로지스틱 모형을 정의하기 위하여 π_{ig}에 대한 로지스틱 항과 결합될 수 있다. 이 모형은 표본크기 $n_{ig}^{o} = round(n_{ig}^{**})$와 표본 실업자 수 $m_{ig} = round(n_{ig}^{o} \times Z_{ig}^{*})$를 입력값으로 갖는 로지스틱 회귀 소프트웨어를 이용하여 실제 표본 자료에 적합될 수 있다. 여기에서 β의 추정값과 $Var(\beta)$의 추정값 $v(\beta)$를 이끌어 낸다. 이때 Z_{ig}의 추정량은 $\pi_{ig} = \text{antilogit}(x_{ig}^{T}\beta)$이 된다. 그러나 이 추정량은 불편성을 만족하지는 않는다. 따라서 편의를 보정한 형태의 추정량은 다음 (3.21)식과 같이 주어질 수 있다.

$$\pi_{ig} = \pi_{ig}\left[1 - \frac{1}{2}(1-\pi_{ig})(1-2\pi_{ig})(x_{ig}^{T} v(\beta)\, x_{ig})\right] \tag{3.21}$$

이때 소지역 g에서 실업자 총계에 대한 추정량은 다음 (3.22)식과 같이 주

어진다.

$$\theta_g = \sum_{i \in g} a_{ig} N^*_{ig} \pi_{ig} \tag{3.22}$$

소지역 g와 h에 대한 모형기반 추정량들 간의 추정 공분산은 다음 (3.23) 식을 통해 계산될 수 있다.

$$c(\theta_g, \theta_h) = \sum_{i \in g} \sum_{j \in h} a_{ig} N^*_{ig} \pi_{ig}(1-\pi_{ig})\left(\boldsymbol{x}_{ig}^T v(\beta)\, \boldsymbol{x}_{jh} \right) \pi_{jh}(1-\pi_{jh}) N^*_{jh} a_{jh} \tag{3.23}$$

3.2.3 LFS 자료를 이용한 적용결과

1995년~'96년과 1998년~'99년의 LFS 자료들을 이용하여 UA지역과 LAD지역에 대한 실업률 추정값들을 계산하였다. 주요 보조변수로써 실업보험 청구자 수를 이용하였다. 이러한 보조변수 및 성별-연령대별 범주의 6개 그룹, 지리적인 권역으로 분류된 12개 그룹과 사회-경제적 분류 집락인 7개 그룹에 대한 척도들이 모형에 포함되었다.

UA지역과 LAD지역에 대한 실업률 추정값은 기존의 직접추정방법보다는 2절에서 언급한 모형기반 추정방법이 상대적으로 변동이 작고 훨씬 안정적으로 나타났다. 모형기반 추정량의 추정오차에 대한 LFS 조사 추정량의 추정오차 비의 평균값은 1995년~'96년 자료에서는 5.5049, 1998년~'99년 자료에서는 5.3851의 값을 나타내며, 모형기반 추정방법이 상대적으로 작은 변동을 나타낸다는 사실을 확인하였다.

3.2.4 추가 연구

2절에서 소개된 것과 같은 합성추정 형태의 모형기반 추정량은 소지역들 간의 변동을 설명할 수 없는 문제점을 안고 있으며, 일반적으로 이러한 방법으로 추정된 추정오차는 과소 추정되는 경향이 있다. 이러한 문제점은 소지역에 대한 랜덤효과를 모형에 반영한 Multi-level 모형을 통해 어느 정도 해소할 수 있으며 이러한 연구가 현재 영국 통계국에서 진행되고 있다. 대상

모형은 $\text{logit}(\pi_{ig}) = \boldsymbol{x}_{ig}^T\boldsymbol{\beta} + u_g$와 같은 모형이다. 여기에서 지역 명시 변수 $\{u_g\}$는 평균이 0이고 분산이 σ_u^2인 확률변수로 가정된다. EBLUP 형태의 성분 추정값 $\pi_{ig} = \text{anti logit}(\boldsymbol{x}_{ig}^T\boldsymbol{\beta} + u_g)$에 기반을 둔 (4)식의 소지역 추정값들은 표본 크기가 큰 UA 및 LAD 지역에서는 LFS 추정값들과 유사하며, 표본 크기가 작은 UA 및 LAD 지역에서는 고정효과를 같는 추정값들과 유사한 경향을 나타낸다. 이러한 추정량이 갖는 실제적인 문제는 추정량의 평균제곱오차(MSE) 계산이 쉽지만은 않다는 데에 있다. 현재 영국 통계국에서는 하나의 절충안으로써 다음과 같은 분산 추정공식을 고려하고 있다.

$$v(\theta_g) = \sum_{i\in g}\sum_{h\in g} a_{ig} N_{ig}^* \pi_{ig}(1-\pi_{ig})\left[\sigma_u^2 + \boldsymbol{x}_{ig}^T v(\boldsymbol{\beta})\, \boldsymbol{x}_{hg}\right]\pi_{hg}(1-\pi_{hg}) N_{hg}^* a_{hg} \quad (3.24)$$

여기에서 σ_u^2은 랜덤효과 모형 적합에서 추정되는 소지역 간의 추정분산을 나타낸다.

3.2.5 결 론

LFS 직접 추정값들은 실업보험 청구자료와 같은 이용 가능한 보조정보를 통해 개선될 수 있다는 사실을 이 논문에서는 보여주고 있다. 이 논문에서 고려된 방법론은 추정값의 정확도를 개선시키기는 하나, 모형에 랜덤효과를 포함시키는 문제와 EBLUP 형태의 추정량들의 평균제곱오차를 추정하는 방법 및 비 추정 방법 등은 여전히 해결되어야 할 문제점으로 남게 된다. 이러한 문제를 해결하기 위한 연구가 현재 영국 통계국 및 Southampton 대학 연구진들에 의해 진행되고 있다.

3.3 캐나다

3.3.1 서 론

(1) 배경 및 목적

캐나다 노동력조사(Labour Force Survey:LFS)는 대규모 노동시장의 변화 양상 및 시의성 있는 노동시장의 정보를 파악하기 위해 2차 세계대전 이후 도입되었고, 주로 주(Province) 지역 및 국가 단위의 고용 및 실업통계를 생산할 목적으로 설계되었다. LFS는 1945년 분기별 조사로 시작하여 1952년 월별 조사로 변경되었고, 1960년부터 캐나다 실업통계를 생산하기 위한 공식 조사로 승인되었다. 그 후 LFS를 통해 노동시장의 다양한 통계를 작성할 수 있도록 표본개편 및 조사방법에 관한 연구가 지속적으로 진행되었고, 현재는 캐나다 노동시장의 세부적인 변화에 관한 정보를 제공할 수 있을 정도로 발전을 거듭하였다. 매월 고용인구와 실업인구 총계 및 실업률에 관한 추정치, 노동인구의 특성(연령, 결혼여부, 교육정도, 가족현황) 등에 관한 공식통계는 LFS를 통해 작성된다.

LFS는 주 지역 및 전국 단위의 고용 및 실업통계 작성 외에 고용보험 경제구역(EIER:Employment Insurance Economic Regions), 센서스 도시지역(CMA:Census Metropolitan Areas) 등과 같은 주 내의 특정 행정구에 대한 통계작성도 가능하도록 표본이 설계 되어있다. 최근에 들어서는 주 지역 내의 센서스 조사구(CD:Census Divisions)와 같은 소지역들에 대해서도 소지역 추정기법을 이용하여 관련 통계지표를 작성하고 있으며, 지방정부의 소지역에 대한 예산 배분 또는 정책 결정 등의 사안에 이러한 소지역 통계들이 이용되고 있다.

LFS 추정치들은 매월 "Labour Force Information"라는 책자를 통해 공표된다. 또한 노동시장의 좀 더 다양한 정보들은 캐나다 통계국의 전자정보 데이터베이스의 일종인 "CANSIM"을 통해 획득할 수 있으며, LFS의 결과로부터 매월 9000 항목 이상의 시계열 자료들이 정기적으로 수정 보완된다. 이외에 노동시장의 중심지표가 되는 다양한 주제에 대한 세부적인 고찰을 다룬 "Labour Force Update"가 1997년부터 계간지로써 출간되고 있으며,

1976년 이래로 최근까지의 방대한 시계열 자료(time series data) 및 횡단면 자료(cross-sectional data)를 포함하고 있는 "Labour Force Historical Review on CD-ROM"이 매년 제작되고 있다.

캐나다 노동력 조사에 의해서 매월 발표되는 통계수치는 자영업, 부업과 전업을 포함한 취업자 총수와 실업자 총수이다. 매월 발표하는 노동시장의 표준지표는 실업률, 취업률과 경제활동 참가율이고 노동력 조사의 주요 정보요소로서 15세 이상 인구의 개인적 특성은 나이, 성별, 혼인상태, 교육정도와 가족사항이다.

취업통계의 추정값들에는 인구학적 특성, 산업과 업종, 정규직과 통상적인 근로시간 등이 포함되어 있으며 설문내용에는 비자발적 부업적 취업, 복수직업 여부와 휴직 등에 대해서 분석할 수 있는 주제들이 포함되어 있다. 특히 1997년 이후에는 근로자들의 노조가입 여부와 임금수준에 대한 정보와 작업장의 근로자 수 및 직업의 정규직 또는 임시직 여부에 대한 정보를 제공하고 있다.

실업통계의 추정값은 인구학적 범주별, 실업기간, 구직활동 전의 활동 및 바로 이전 직장에서 이직한 이유 등에 대한 정보를 제공하고 있다. 노동력 조사에 의해서 발표되는 통계는 국가 단위와 주 단위 추정값이 핵심적인 내용이지만 경제구역(ER : Economic Region)과 센서스 도시지역(CMA ; Census Metropolitan Area)과 같은 소지역 단위에 대한 노동력 상태의 추정값을 제공하고 있다.

(2) 노동력 상태 결정과정

취업과 실업의 개념은 생산의 요소로서 노동력 공급이론을 근거로 정의하였으며 생산은 국민계정 체계(SNA : the System of National Acconts)에서 언급한 것과 같이 상품과 서비스로 정의되는 개념이므로 작업의 목적이나

성질에서는 보수를 받는 근로활동과 조금도 차이가 없는 무보수의 가사노동이나 자원봉사활동은 근로시간으로 계산하지 않고 있다.

노동력 공급의 측정단위는 개별적인 근로시간이지만 조사에서 모집단의 개별적인 구성원들의 구분은 취업, 실업, 비경제활동 인구로 분류되어야 한다. 조사기간 중에 보수 근로 중인 사람은 근로시간에 상관없이 취업자로 구분하고 근로시간과 무관하게 노동시장에서 구직 행위가 있을 경우에는 실업자로 구분한다. 나머지 인구는 현재 일을 하고 있지 않거나 또는 노동시장에서 구직 활동하지 않는 경우로 비경제활동 인구로 정의한다.

통계조사에서 적용하는 취업자와 실업자의 정의와 개념은 국제노동기구(ILO : International Labor Organization)의 기준에 준거하고 있다.

(가) 취업자(Employment)

조사대상 기간 중에 직업이 있거나 개인 사업에서 일을 하는 사람을 말하며 자기에게 직접적인 소득이 없을 지라도 가구단위로 운영되는 농장, 사업체 또는 전문적인 기관에서 일하는 경우도 포함된다. 또한 직업이나 사업체가 있을지라도 일시적인 병이나, 휴가, 노동쟁의 등의 이유로 일을 하지 못하는 일시적인 휴직자도 포함한다.

(나) 실업자(Unemployment)

이용되지 못하는 공급된 노동력으로 실업자를 정의하고 있으며 실업자의 구분은 구직행위와 근로행위의 준비여부를 기준으로 하고 있으나 구직행위는 가구조사에서 구체적이고 지속적인 의사 표명을 전제로 하고 있다. 실업자는 조사 대상 기간 중 다음 3가지 항목 중 하나에 해당하는 사람으로 정의될 수 있다.

1) 현재 휴직 중이지만 근로 활동이 가능하고 복직을 기대할 수 있는 경우

2) 일을 하지 않고 있으나 구직 활동을 하고 있으며 일이 주어지면 바로

시작할 수 있는 경우

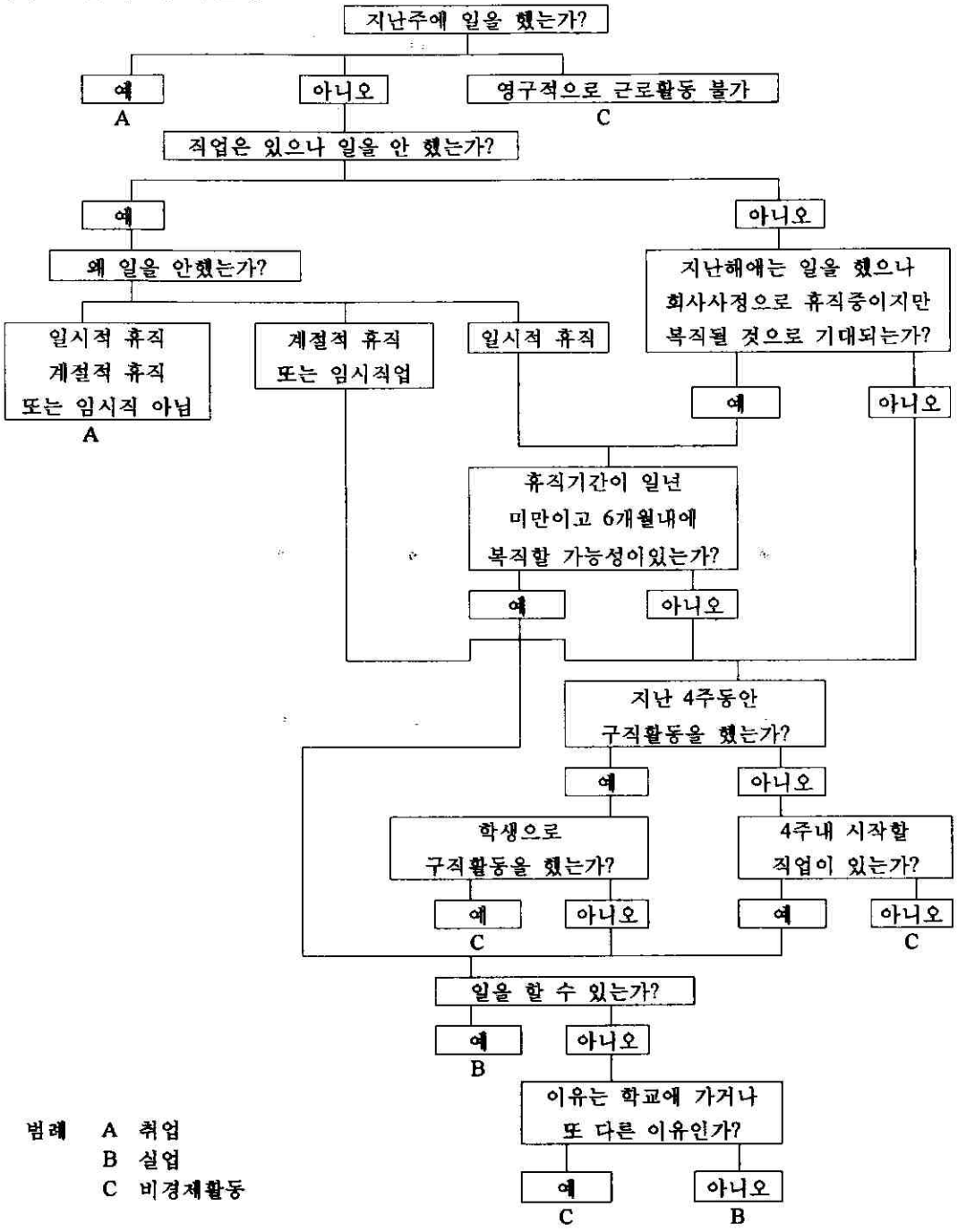
3) 조사대상 기간으로부터 4주 이내에 새로운 일을 할 직업이 주어졌으며 바로 일을 할 수 있는 경우

현재 학교에 다니는 학생이 전업적인 일을 찾고 있을 지라도 구직활동으로 간주하지 않으며 여름방학동안의 근로 활동이나 구직 활동 역시 실업자의 분류로 고려하지 않고 있다.

(다) 비경제활동인구

노동시장에서 노동력을 제공하거나 참여할 의사가 없거나 가능성이 없는 사람을 말하며 이들은 실업자도 아니고 취업자도 아닌 상태의 사람들이다. 여기에는 순수하게 가사만 돌보는 사람, 학생, 투자배당금이나 연금 등을 받아서 생활하는 사람들이 포함된다.

■ 노동력 상태결정



LFS는 캐나다 인구의 약 98%에 해당하는 인구를 목표모집단(target

population)으로 설정하였다. 캐나다 북서부지역, 인디언 보호구역, 왕실 소유지, 수감자 및 직업군인에 해당하는 약 2%는 LFS의 조사대상에서 제외된다. 1998년 12월 현재 LFS의 표본크기는 52,350가구이다. 1970년 표본개편을 통해 LFS의 표본크기는 이전의 35,000가구에서 50,000가구로 증편된 후, 1980년 47,000가구로 감소되었다가 1989년 주 지역 내의 행정구역인 EIER 지역의 통계작성에 신뢰성을 확보하기 위해 63,000조사가구로 표본설계가 대폭 개편되었으며, 1993년에 약 59,000 조사가구로 다시 감소된 후 1995년부터 52,350 조사가구(EIER 지역에 16,500 조사가구, 나머지 지역에 35,850 조사가구)로 조정되어 현재에 이르고 있다.

LFS의 표본교체 체계(sample rotation)는 일종의 연동교체표본설계(rotating panel sample design)를 따른다. 표본가구는 6개의 부차표본(sub-sample)로 분리되어 6개월간 관리되며, 매월 1/6의 표본이 새로운 표본으로 대체되는 형식이다. 캐나다의 LFS는 매월 15일이 포함되는 주중에 실시되며, 80명의 조사전문인력을 포함하여 약 850명의 조사인력이 투입되어 조사가 이루어지고, 조사결과는 5개의 지방사무소(RO: Regional Office)에서 각각 취합되어 중앙으로 이관된다. 첫 달의 조사는 방문면접 형식을 취하며, 이 후 연속되는 다섯달 은 전화조사를 통해 조사가 이루어진다. 조사자는 휴대용 컴퓨터를 이용하여 직접 설문 항목의 결과를 입력하는 일종의 컴퓨터 보조 면접(CAI: Computer Assisted Interviewing) 방식을 이용한다. 조사결과는 조사완료 시점에서 정확히 13일 후에 공표된다.

3.3.2 층화 및 추출단위 구성

캐나다의 주(Province) 지역들은 지리적인 경계에 의해 여러 개의 경제구역(ER:Economic Regions)들로 분할되며, 현재 캐나다 ER 지역은 총 72개가 존재한다. LFS에서는 1960년대 이후로 이러한 ER 지역들을 캐나다 노동력

조사의 1차 층(primary strata)으로 이용해오고 있다. 주 지역 및 전국 단위의 통계 작성에 ER 지역들이 이용된다.

초기의 LFS에서는 ER 지역이 표본설계 시 반영되었던 주 지역 내의 유일한 행정구역이었고 노동력 조사의 관심은 이러한 ER 지역에 집중되었으나, 1989년부터 HRDC(Human Resources Development Canada)의 자금지원에 의해 16,500개의 표본조사가구가 LFS에 추가되면서 EIER 지역 에 대한 노동력 조사도 추가적으로 가능하게 되었다. 따라서 현재의 노동력조사에서는 ER 지역과 EIER 지역 모두가 표본설계 시 층화에 반영되며, 추가표본은 주로 EIER 지역의 추정치의 신뢰도를 확보하기 위해 할당된다. 여기에서 ER 지역과 EIER 지역은 서로 조사 목적이 상이한 지역들이며 133개 지역이 서로 중복되어 조사된다. 한편 LFS 층화 시 반영되는 행정구역으로 CMA 지역을 들 수 있다. CMA 지역은 인구 100,000명 이상인 지역들로써 현재의 CMA 지역은 EIER 지역과 정확히 일치한다.

대영역 내에서의 세부적인 층화는 지리적인 구분에 관계없이 집락화 알고리즘에 의해 이루어진다. 그룹들 간의 가중 제곱합을 최소화하는 층화변수들을 이용하여 가능한한 동질적인 층으로 분할되며, 세부적인 알고리즘은 Drew et al.(1985)과 Singh et al.(1990)에서 참조할 수 있다. 층화 알고리즘에 이용되는 층화변수들은 다음과 같다. 이 층화변수들은 1991년 센서스 자료를 이용하여 선정되었으며, 각각의 층화변수들은 전체인구의 2%이상을 설명할 수 변수들로 선정되었다.

- o 농업부문 종사자 수
- o 임업, 어업부문 종사자 수
- o 광업부문 종사자 수
- o 제조업(소비재분야) 종사자 수
- o 제조업(고무, 플라스틱, 가죽분야) 종사자 수

o 제조업(섬유, 의류 분야) 종사자 수

o 제조업(가구, 펄프, 제지, 인쇄, 목재분야) 종사자 수

o 제조업(금속, 광업분야) 종사자 수

o 제조업(석유화학, 화학분야) 종사자 수

o 운수업 부문 종사자 수

o 건설업 부문 종사자 수

o 서비스업(상업분야) 종사자 수

o 서비스업(금융분야) 종사자 수

o 서비스업(개인/사업분야) 종사자 수

o 서비스업(정부분야) 종사자 수

o 종사인원 총계

o 총 소득

o 15세 이상 인구

o 15-24세 인구

o 55세 이상 인구

o 1인 거주 가구 수

o 2인 거주 가구 수

o 개인 소유 가구 수

o 총 임차료

o 고졸학력 인구

o 영어를 모국어로 하는 인구

o 프랑스어를 모국어로 하는 인구

o 영어/프랑스어 이외의 언어를 모국어로 하는 인구

LFS 추출틀은 농촌 지역, 인구 50,000명 이상의 대도시 지역과 소도시 지역의 3가지 유형의 지역들로 구분된다. 각 지역에 대한 층화는 다음과 같

은 방법으로 이루어진다.

농촌지역에서 층화는 EI 지역과 EIER 지역의 교집합 내에 있는 2~3개의 CD(Census Division) 지역들을 함께 묶어 지리적 층으로 구성하였다.

캐나다 내의 대도시 지역인 17개의 CMA 지역들에 대해서는 충분한 수의 아파트들이 있기 때문에 각각의 CMA 지역들에 대해 독립적인 아파트 추출틀을 작성하며, 아파트 추출틀을 제외한 나머지 지역에 대해서는 일종의 지역 추출틀(area frame)을 형성하였다. 또한 $100,000 이상의 평균소득을 갖는 고소득 지역들은 독립된 층으로 구성하여 고소득 가구들에 대한 대표성을 제고하였고, 부수적으로 소득관련 조사 및 소득관련 정보 수집이 용이하도록 하였다. 이러한 부수적인 정보는 고소득층에 대한 무응답의 경향을 분석하고자 하는 경우에도 유용하게 이용될 수 있다. 아파트 추출틀과 고소득 층을 제외한 나머지 지역들은 SNF(Street Network File) 지역으로 구분하였다. 최종적으로 마지막 층에는 적어도 48가구의 표본이 배정되도록 하였다.

LFS 표본설계에서 대규모 CMA 지역들에 대해서는 아파트 추출틀을 사용해오고 있다. 현재 18개 CMA 지역에서 표본 추출틀로써 이 목록이 이용되며, 각 CMA 지역에서 새로운 아파트가 건설되면 바로 표본목록에 추가된다. 또한 7개의 CMA 지역에 대해 평균소득이 $20,000 미만의 저소득 아파트 단지를 파악하여 저소득 아파트 층을 구성하여, 고소득 아파트 층과 더불어 소득관련 정보를 획득한다. 캐나다의 각 CMA 지역에 대한 아파트 추출틀 층화 현황은 다음 <표3.8>에 주어졌다.

<표 3.8> 아파트 추출틀 층화

CMA	지리적 층	층의 총수	CMA	지리적 층	층의 총수
Halifax	2	2	London	1	2
Quebec	2	2	Windsor	1	2
Montreal*	4	9	Winnipeg*	1	6
Ottawa-Hull*	3	6	Saskatoon	1	1
Oshawa	1	2	Calgary*	1	3
Toronto*	6	16	Edmonton*	1	3
Hamilton	2	4	Vancouver*	4	6
St. Catharines	1	1	Victoria	1	1
Kitchener	2	2	Total	34	68

(주) ① "*" 는 저소득 아파트 층을 갖는 CMA 지역을 표시함

② "층의 총수"는 저소득 아파트 층을 포함한 수임

대부분의 소도시에서는 EA(enumeration area) 지역을 층화단위로 이용하였다. 조사 비용을 절약하기 위해 각 층에서 직접 표본가구를 추출하지 않고, 우선 각 층을 여러 개의 집락으로 구분하고 추출된 집락 내에서 표본가구들을 추출하였다. 통상적으로 농촌지역에서는 EA 지역이 집락으로 사용되었고, 도시지역에서는 좀 더 다양한 형태의 집락들이 이용되었다. 다음 <표 3.9>는 LFS 표본설계에 이용된 일 단계 단위(first-stage unit)의 유형들을 요약한 것이다. 여기에서 "추출가구 수"는 LFS에서 조사되는 가구 수를 나타낸다. 집락과 표본가구에 대한 추출방법은 3장에서 논의하기로 한다.

<표 3.9> 일단계 단위(first-stage unit), 단위당 가구 수와 추출가구 수

지 역	추출단위	단위당 가구수	추출가구수
Toronto, Montreal, Vancouver	cluster	200-250	6
Other cities	cluster	150-200	8
Apartment frame	apartment	varies	5
Most rural areas and non-SNF prarts of cities	EA	300	10

3.3.3 표본배정, 추출, 순환(Sample Allocation, Selection and Rotation)

LFS의 표본설계는 캐나다 전체, 주(Province) 지역, EIER 지역, CMA 지역, ER 지역에 대해 각각 다음과 같은 실업자 추정값의 목표 CV값이 만족되도록 설계되었다. 캐나다 전체의 실업자 추정치의 목표 CV 는 약 2%이내, 주(Province) 지역의 CV 값은 약 4~7%선에서 관리되도록 하였다. EIER 지역과 CMA 지역은 분기 실업자 추정치의 CV 값이 15% 이내가 되도록 하였고, 여기에서 하나의 EIER 지역에 배당되는 최소 표본크기는 매월 600가구가 배정된다. ER 지역에 대한 분기별 목표 CV 값은 25% 이내가 되기를 기대하고 있다. 캐나다의 각 주 내의 ER 지역, EIER 지역과 CMA 지역의 현황은 다음 <표3.10>과 같다.

LFS의 표본크기는 총 52,350가구로써 HRDC의 자금지원에 의해 추가된 16,500 표본가구를 제외한 35,850 표본가구는 전국 및 주(Province) 지역의 최적 추정을 위해 배정되며, 이 표본을 핵심표본(core sample)이라고 지칭한다. 핵심표본은 각 주의 실업률 추정의 목표정도를 만족하도록 배정된다. 주 내에서 각각의 ER 지역에 대한 핵심표본의 배정은 총 가구 수에 비례하여 배정하며, 최소한 200~300 표본가구가 배정되도록 하였다. 또한 16,500 가구의 추가표본은 핵심표본을 보충하여 EIER 지역의 추정의 정확도를 높이기 위해 추가로 배정된다. 우선 핵심표본 만으로 EIER 지역에 대한 CV 값을 계산 한 후, CV 값이 큰 EIER 지역에 대해 추가 표본을 배정하는 형식을

취하며, EIER 지역에 대해서는 최소한 600 가구가 배정되도록 하였다.

<표 3.10> ER, EIER, CMA 지역 현황

주(Province)	ERs	EIERs	CMAs
Newfoundland	4	3	1
Prince Edward I	1	1	0
Nova Scotia	5	5	1
New Brunswick	5	4	1
Quebec	16	13	6*
Ontario	11	18	10*
Manitoba	8	3	1
Saskatchewan	6	4	2
Alberta	8	4	2
British Columbia	8	6	2
Canada	72	61	25*

(주) Ottawa-Hull CMA 는 Ontario와 Quebec 양쪽에서 카운트 됨.

LFS에서 표본 추출은 대부분의 지역에서 이 단계 추출(two-stage sampling) 만을 이용한다. 일 단계 추출은 지역 추출(area sample)이고, 이 단계 추출은 일 단계 추출 지역들에 대해 거주가구 목록을 작성하여 이 목록으로부터 표본가구를 추출한다.

농촌지역의 경우 일 단계 추출단위는 EA 지역, 이 단계 추출단위는 가구 단위가 된다. 일 단계의 지역 추출은 확률비례계통추출법, 이 단계의 가구 추출은 계통추출법이 이용된다. 아파트 추출틀을 이용하지 않는 주요 도시지역의 경우 일 단계 추출단위는 EA 지역, 이 단계 추출단위는 가구 단위가 된다. 일 단계의 지역 추출은 Rao-Hartley-Cochran(RHC)의 랜덤그룹법이 이용되며, 이 단계의 가구 추출은 계통추출법이 이용된다. 아파트 추출틀을 이용하지 않는 기타 도시 지역들도 몇몇 특별한 경우를 제외하고는 주요 도시지역과 유사한 방법을 취하고 있다. 랜덤그룹법은 인구유입 및 유출양상을 LFS의 표본크기에 반영시킨 방법으로써 시기에 따라 유동성이 큰 대도시

지역의 인구변화를 표본크기에 반영시킨 방법이다. 아파트 추출틀을 이용하는 주요 도시지역의 경우 일 단계 추출단위는 아파트 단지, 이 단계 추출단위는 가구 단위가 된다. 일 단계 아파트 단지에 대한 추출은 확률비례계통추출법, 이 단계 가구단위의 추출은 계통추출법이 이용된다.

LFS에서는 표본의 일부가 매달 새로운 표본으로 교체된다. 다 단계 표본설계의 매 단계에서 표본 단위의 교체가 이루어지며, 표본추출의 최종 단위인 가구는 6개월마다 교체된다. 조사자의 업무 부담과 응답가구가 오랜 기간 동안 표본으로 조사되는 동안 발생할 수 있는 무응답에 대한 가능성을 최소화하기 위해 매 달 1/6의 표본이 대체된다. 따라서 하나의 집락에 포함된 표본가구는 연속적으로 6개월 간 조사된 후 표본에서 완전히 삭제되고 새로운 표본으로 대체된다.

3.3.4 특별조사와 보충조사

캐나다 노동력조사 외에 추가적인 많은 조사가 LFS 추출틀 또는 LFS 표본을 이용하여 실시되며, 이러한 조사들은 캐나다 정부 부처의 자금 지원에 의해 시행된다. 다음 <표3.11>은 LFS 연동교체표본 또는 LFS 추출틀을 이용한 특별조사 및 보충조사 현황을 요약한 것이다.

<표 3.11> 특별조사 및 보충조사 현황(1998년 현재)

Survey	조사기간	Survey	조사기간
Canadian Travel Survey	1-12월(매월조사)	Homeowner Repair and Renovation Survey	3월
Employment Insurance Coverage	1월	Survey of Consumer Finances	4월
Survey of Household Spending	1월-3월	Cultural Capital Survey	4월
Survey of Labour and Income Dynamics	1월, 5월	National Population Health Survey	2, 6, 8, 11월
Adult Education and Training Survey	1월	National Longitudinal Survey of Children	11월
Resident Telephon Services Survey	2, 5, 8, 11월	Survey of Work Arrangements	11월
Survey of Household Energy Use	2월		

SCF(Survey of Consumer Finances)는 일년에 한번 시행되는 소비자 재정에 관한 조사로써 보통 4월에 실시된다. 모든 가구들이 4개의 순환 그룹에 배정되어 LFS 조사에 추가된다. 4월의 LFS 조사에 앞서 각 가정에 우편으로 설문지가 배달되고 LFS 조사 기간 동안 회수된다. SCF에서 조사하는 주요 정보는 소비자의 평균 소득과 세금 공제 전과 공제 후의 소득관련 정보들로써, 이러한 결과들은 저소득 기준점 결정 등과 같은 소득 관련 측도들로 이용된다.

SHS(Survey of Household Spending)는 일년에 한번 실시되는 가계비 지출 및 식료품 지출 조사로써 보통 1월~3월에 걸쳐 시행되며, 소비자 가격지표 산정의 정보로써 이용된다. SHS 조사는 LFS 표본을 포함하는 집락들에서 표본가구를 추출하나 표본가구들은 LFS 조사와는 별도로 조사된다.

SLID(Survey of Labour and Income Dynamics)는 노동력과 소득 변천과정 파악을 조사의 목적으로 일년에 2번, 1월과 5월에 조사가 이루어지며 LFS와 병행하여 시행된다. 조사 결과는 저소득 층의 유입, 유출 동향, 노동시장의 변화, 가족변화와 경제적인 복지와의 상관관계 등을 분석하기 위해

이용된다.

NPHS(National Population Health Survey)는 일종의 국민 건강조사로써 분기별로 실시되며 국민 건강의 계절적 요인을 파악하기 위해 NPHS 표본을 2월, 6월, 8월과 11월조사에 각각 1/4씩 배분한다. NPHS 조사는 주 정부의 자금지원에 의해 실시되며, LFS 조사와 병행하여 실시되지는 않는다. 초기에 선택된 가구의 구성원은 2년에 한번 심층 면접을 받으며 20년 동안 지속적으로 관리된다. 기초적인 건강정보는 거주 가구의 모든 구성원들에 대해서 취합되며 이러한 시계열 자료는 횡단면 추정 목적에 이용된다.

NLSCY(National Longitudinal Survey of Children and Youth)는 아동에 대해 유아기부터 성인기까지의 발달 과정을 모니터하는 조사로써 일년에 한번 실시되는 일종의 시계열적 장기조사에 해당한다. LFS 조사가구의 약 30%만이 대상연령에 있는 아동을 포함하는 관계로 NPHS의 추가표본이 조사에 이용되며, 조사방법은 NPHS 조사와 유사하다.

3.3.5 가중치와 추정

LFS의 가중치는 다음의 세가지 요인들에 기인하여 작성된다. 표본설계를 반영하는 일종의 설계 가중치, 무응답 가구를 보정하는 일종의 무응답 보정 가중치와 모집단 총계에 표본 추정치를 일치시키는 일종의 사후 층화 가중치(g-factor)의 3가지 요인들에 의해 LFS의 최종 가중치가 결정된다. LFS 조사 추정치는 LFS 표본이 확률표본이기 때문에 추정치의 표본오차를 추정하여 신뢰도를 판단할 수 있다. 표본설계에 계획되지 않은 관심지역들에 대한 추정문제는 소지역 추정기법을 도입하여 추정량의 신뢰도를 확보하였다. EIER 지역들이 여기에 해당된다.

LFS에서는 가구단위에 대해 계통추출이 이루어지는 마지막 단계를 제외하고는 모든 단계에서 확률비례추출법으로 표본이 추출되는 층화 다단계

추출법이 이용된다. 설계가중치는 이러한 복합표본설계에서 가장 기본적인 가중치로써 각 추출단위에 대해서 추출률의 역수로 결정된다. 예를 들어 층에 대한 설계가중치는 $R_h = N_h/n_h$와 같이 표현될 수 있다. 여기에서 n_h는 층 h에 대한 표본가구 수, N_h는 표본설계 시 층 h에 있는 총 가구 수를 나타낸다.

이 단계 추출의 경우는 다음과 같이 설명될 수 있다. n^*_{hj}를 층 h에 있는 j번째의 일 단계 추출단위(FSU: First Stage Unit)라고 하자. 추출해야 할 FSU의 수는 $n_{1h} = n_h/n^*_{hj}$ 로 주어진다. h층에서 j번째 FSU에 있는 가구들의 수를 N_{hj}라 하면 j번째의 FSU에 대한 추출율은 $R_{hj} = N_{hj}/n^*_{hj}$로 주어지며, 이때 j번째 FSU에 대한 일 단계 산입확률(inclusion probability)은 $\pi_{1hj} = \frac{n_{1h}}{\sum_{j\in h} R_{hj}} R_{hj}$로 나타낼 수 있다. j번째 FSU가 주어진 상태에서 k번째 가구가 선택될 조건부 산입확률은 $\pi_{k/j} = \frac{n^*_{hj}}{N_{hj}} = \frac{1}{R_{hj}}$로 주어지며, h번째 층에서 k번째 가구에 대한 산입확률은 $\pi_{hk} = \pi_{1hj} \cdot \pi_{k/j} = \frac{n_{1h}}{\sum_{j\in h} R_{hj}} R_{hj} \frac{1}{R_{hj}} = \frac{n_{1h}}{\sum_{j\in h} R_{hj}}$로 계산될 수 있다. 여기에서 $\sum_{j\in h} R_{hj} = \sum_{j\in h} \frac{N_{hj}}{n^*_{hj}} = \frac{n_{1h}}{n_h} \sum_{j\in h} N_{hj} = n_{1h} R_h$ 인 관계가 성립하며, 결국 산입확률은 $1/R_h$과 같게된다. 참고적으로 LFS에서는 기본적으로 각 층내에서 동일한 설계 가중치(R_h)를 이용한다.

LFS와 같은 연속조사에서는 고정된 추출률을 유지할 경우, 인구 증가 및 인구 유입 등으로 시간이 지남에 따라 표본 수도 증가하게 된다. 급속한 인구 증가가 발생하는 특정지역의 표본 수는 급격히 증가되어 조사원의 업무 부담을 가중시키고 조사의 질을 떨어뜨릴 가능성이 있다. 이러한 문제점을

해결하기 위해 LFS 표본 설계에서는 이러한 집락들에 대해서 표본을 부차추출(subsampling)하여 전체 표본크기를 일정하게 유지한 후 집락 부차가중치(cluster subweight)를 산출하여 LFS 설계가중치에 추가하여 사용하였다. 구체적으로 SC(Subclustering)방법, SRC(Self- Representing Cluster)방법, CSS(Cluster Subsampling)방법 등을 이용한다.

SC 방법은 인구유입 등으로 인해 원래의 집락의 크기가 약 3배를 초과할 때 이러한 집락들을 여러개의 작은 집락으로 분할하여 표본 가구를 추출하는 방법이다. 분할된 작은 집락들의 표본 수를 n_{2hj}라 하자. N_{hji}를 새로운 집락의 크기, n_{hji}를 새로운 집락의 표본크기, R_{hji}를 새로운 집락들에 대한 추출률이라 할때, 새로운 부차집락들로부터 집락추출률 $R^*_{hj}=\sum_{i\in j}\frac{R_{hji}}{n_{2hj}}$를 얻을 수 있다. 이때 집락 부차가중치는 $K=R^*_{hj}/R_{hj}$로 주어지며, 설계가중치는 원래의 가중치에 이러한 부차가중치를 곱하여 계산된다.

SRC 방법은 임의의 층에서 증가하는 거주단위들의 특성이 나머지 단위들의 특성과 구별되거나 집락의 규모가 층 규모의 20%를 초과할 경우, 해당 집락을 하나의 층으로 재분류하여 가중치를 결정하고, 나머지 집락들에 대해서는 가중치를 재 보정해 주는 방법이다. 이렇게 생성된 새로운 층을 (hj)라 하자. 이러한 층내에서 새로운 집락들이 구성되어 표본들이 추출된다. $N^N_{(hj)}$을 새로운 층의 크기, $n^N_{(hj)}$을 새로운 층의 표본크기라 할때, 층 추출률은 $R^N_{(hj)}=N^N_{(hj)}/n^N_{(hj)}$으로 주어지며, 이러한 새로운 층으로부터 추출된 가구들은 가중치 $K=R^N_{(hj)}/R_h$가 배정된다. 나머지 집락들의 가구들에 대한 가중치는 다음과 같은 방법으로 보정된다. 하나의 층에서 6개의 집락들이 추출되었다고 하자. 층 내에서 1개의 성장집락이 발생하여 새로운 층으로 재 분류 되었다면 나머지 5개의 집락의 가구들에 대한 가중치는 보정되어야만 한다. $N^R_h=N_h-N^N_{(hj)}$를 새로운 층을 제외한 층의 크기, $n^R_h=n_h-n_{hj}$을 표본의 크

기라 할때, 층에 대한 추출률은 $R_h^R = N_h^R / n_h^R$로 계산되며, 이때 집락 부차가중치는 $K = R_h^R / R_h$로 결정된다.

CSS 방법은 집락단위가 부차표본으로 추출되는 경우에 이용된다. 임의의 집락이 추출률 R_{hj}로 추출되고, 부차추출이 추출률 R_{hj}^*로 추출되었다면, 이때 집락 부차가중치로써 $K = R_{hj}^* / R_{hj}$를 이용하는 방법이다.

표본추출의 마지막 단계는 계통추출이 이용된다. 가구에 대한 추출률은 일관되게 적용되기 때문에 인구 증가로 인한 표본가구수의 증가는 조사규모 및 조사비용의 증가를 초래한다. 이러한 조사비용을 통제하기 위해 표본 수 안정화 작업이 수행된다. 표본 수 안정화 작업은 전체 표본가구 수를 적절한 수준에서 유지하기 위해 초과 표본들을 랜덤하게 제거하는 작업을 말한다. 이러한 과정에서 가구단위의 산입확률(inclusion probability)은 당연히 바뀌게 된다. 현 표본설계에서는 같은 EIER 지역에 속해 있고 같은 순환그룹에 포함되어 있는 모든 가구단위들을 안정화 지역(stabilization area)으로 정의하고, 각 안정화 지역(a)에 대해서 표본크기가 결정된다. 안정화 지역 a의 표본크기를 b_a라고 나타내자. 안정화 지역에서 안정화작업을 거치지 않은 표본크기를 n_a라 할 때, n_a가 b_a를 초과한다면 $n_a - b_a$만큼의 표본가구가 랜덤하게 제거된다. LFS에서는 임의의 집락이 부차추출 되었을 경우에는 이 집락을 안정화 작업에서 제외시키기 때문에 이러한 집락에 포함된 가구단위들은 안정화 가중치(stabilization weight)의 영향을 받지 않는다. 이러한 집락을 제외시킨 안정화 지역 a의 가구단위의 총계를 c_a라 할때, 지역 a에 있는 조사가구의 안정화 가중치는 $s_a = (n_a - c_a)/(b_a - c_a)$이 이용된다.

통계조사에서 무응답은 항목 무응답(item nonresponse)과 단위 무응답(unit nonresponse)의 두 가지 유형으로 분류된다. LFS에서는 항목 무응답은 대체법(imputation)으로, 단위 무응답은 전체적인 가중치 조정을 통해 처

리하고 있다. 항목 무응답은 지리적으로 또는 인구 통계적으로 유사한 특성을 갖는 응답자의 응답패턴을 이용하여 대체된다. LFS에서는 단위 무응답을 처리하기 위한 무응답 층을 " 같은 EIER 지역에 속하고, 같은 유형의 지역적 특성을 가지며, 같은 표본순환그룹 내에 있는 가구들"로써 정의한다. 무응답 층을 올 바르게 구성하였을 경우 동일한 무응답 층에 속한 응답 가구와 무응답 가구는 서로 비슷한 속성을 갖기 때문에 응답 가구가 무응답 가구를 대표한다고 가정할 수 있게 된다. 단위 무응답에 대한 보정은 설계 가중치에 보정요인 $f_b = \sum_{k=1}^{n} \pi_k^{-1} / \sum_{k=1}^{r} \pi_k^{-1}$ 을 곱하여 계산한다. 여기에서 π_k^{-1} 은 각 표본가구에 부여된 설계 가중치, n 은 무응답 층 b에 있는 표본가구 수, r 은 응답가구 수를 나타낸다.

LFS에서의 최종 가중치는 특별한 경우를 제외하고는 설계 가중치와 보조정보로부터 얻게되는 사후 가중치(g-factor)의 곱으로 계산되며, 여기에서 사후 가중치 계산은 일반적인 회귀추정방법이 이용된다. 자세한 추정절차는 Lemaitre and Dufour (1987)에 소개되어 있다. 먼저 다음과 같은 기호를 정의하자.

p = 1, 2, ···, 10 : 주 지역을 나타내는 기호,

u= 1, 2, ···, U : p번째 주 내에 있는 EIER 지역,

f = 1, 2, ···, F : u번째 EIER 지역 내에 있는 추출틀의 형태,

h = 1, 2, ···, H : 추출틀 f 내에 있는 층,

r = 1, 2, ···, 6 : h번째 층 내에 있는 순환그룹,

j = 1, 2, ···, J : 순환그룹 r 의 집락,

k = 1, 2, ···, K : 집락 j에서의 가구,

i = 1, 2, ···, c_k : k번째 가구 내에 있는 구성원,

SC(subclustering) 방법은 우선 해당 집락을 여러개의 작은 집락들로 재구성한 후 표본 추출을 위한 집락들을 선정하고 전체 표본 수를 참고하여

선정된 집락 내에서 표본가구를 추출한다. 원래의 집락 추출율이 $R_{pu/h\cdot j}$, 해당 집락 추출율이 $R^{*}_{pu/h\cdot j}$라 하면 해당 집락의 부차 가중치는 $c_{pu/h\cdot j}=R^{*}_{pu/h\cdot j}/R_{pu/h\cdot j}$이다.

SRC(self-representing cluster) 방법은 층 내에서 성장 집락을 분리하여 새로운 층(h)으로 구성하고 h층 내에서 여러개의 집락들을 재 구성한 후 표본가구를 추출하는 방법이다. 원래의 층 추출율이 $R_{pu/h}$, 새로 형성된 층의 추출율이 $R^{*}_{pu/h}$라 하면 이때 새로 형성된 층에서 가구단위들에 배정되는 집락 부차가중치(cluster subweight)는 $c_{pu/h}=R^{*}_{pu/h}/R_{pu/h}$이다. 성장집락을 제외한 나머지 집락의 가구단위에 대해서는 보정된 집락 부차가중치 $c_{pu/h}=R^{R}_{pu/h}/R_{pu/h}$를 적용한다. 여기에서 $R^{R}_{pu/h}$은 나머지 층의 추출율을 나타낸다.

CSS(cluster Subsampling) 방법은 추출된 가구단위들이 부차추출되고 이러한 부차추출된 가구단위들만 조사하는 방법이다. 원래의 집락 추출율이 $R_{pu/h\cdot j}$이고 부차추출하기 위한 해당 집락 추출률이 $R^{*}_{pu/h\cdot j}$라면 이때 집락 부차가중치는 $c_{pu/h\cdot j}=R^{*}_{pu/h\cdot j}/R_{pu/h\cdot j}$이다.

안정화 가중치(stabilization weight)는 앞서 언급되었던 안정화지역(stabilization area) 내에서만 적용된다. 각 안정화 지역 내에서의 표본크기를 $b_{pu\cdot\cdot r}$, 실제 추출된 표본크기를 $n_{pu\cdot\cdot r}$, 안정화 지역에서 CSS 방법으로 추출된 표본의 크기를 $c_{pu\cdot\cdot r}$라 할때, 안정화 가중치는 $s_{pu\cdot\cdot r}=\frac{n_{pu\cdot\cdot r}-c_{pu\cdot\cdot r}}{b_{pu\cdot\cdot r}-c_{pu\cdot\cdot r}}$ 로 계산된다.

이상의 가중치들을 이용하여 조사가구에 대한 설계가중치를 산출하며 LFS에서는 다음과 같은 설계가중치를 고려한다.

$$\pi^{-1}_{pu/h/rik}=w_{pu/h}\times c_{pu/h\cdot j}\times s_{pu\cdot\cdot r},$$

여기에서 w_{psfh}는 같은 층 내에 있는 모든 가구단위들에 대해서 동일한 가중치를 배정했던 표본설계 당시의 가중치를 나타낸다. 표기상의 편의를 위해 앞으로 $\pi^{-1}_{psfkrjk}$를 π_k^{-1}로 나타내기로 한다.

LFS에서는 무응답 층에 대해서 무응답 보정을 실시하며, 보정 가중치로써 $f_{psf\cdot r}=\sum_{k\in s}\pi_k^{-1}/\sum_{k\in r}\pi_k^{-1}$를 이용한다. 여기에서 분자의 s에 대한 합은 무응답 층에 있는 모든 가구들에 대한 합을 나타내며, 분모의 r에 대한 합은 무응답 층에 있는 모든 응답가구들에 대한 합을 나타낸다. 같은 무응답 층에 속해 있는 모든 가구단위들은 동일한 무응답 가중치를 갖는다.

무응답 보정요소가 추가될 경우 부차 가중치는 $a_k=f_{psf\cdot\cdot r}\times\pi_k^{-1}$와 같이 설계가중치와 무응답 보정가중치의 곱으로 표현된다. 즉 같은 조사가구 내의 모든 구성원들은 동일한 부차가중치를 갖는다.

위에서 언급한 부차가중치를 이용하여 고용 인구 Y에 대한 총계 추정값을 산출해 보자. 모집단에서 고용인구의 총계를 $t_y=\sum_U y_i$라고 하자. 여기에서 U에 대한 합은 모집단에서 관심영역 내에 있는 모든 구성원들의 합을 나타내며, y_i는 조사대상자가 고용일 경우 1, 아닐 경우 0의 값을 갖는다. 이때 표본조사에 의한 총계 추정치는 부차가중치에 의존하며 $\hat{t}_{ya}=\sum_s y_i a_i$와 같이 표현될 수 있다. 여기에서 s에 대한 합은 표본으로 추출된 조사 대상자들에 대한 합을 나타내고, a_i는 부차가중치를 의미한다. 위의 t_y와 $\hat{t}_{ya}$는 각각 다음과 같이 다시 표현할 수 있다.

$$t_y=\sum_{k=1}^{N}\sum_{i=1}^{c_k} y_i=\sum_{k=1}^{N} y_k \quad , \quad \hat{t}_{ya}=\sum_{k=1}^{n} a_k\sum_{i=1}^{c_k} y_i=\sum_{k=1}^{n} y_k a_k \ ,$$

여기에서 c_k는 k번째 조사가구의 구성원의 수, N은 모집단의 가구 수, n은 표본 가구 수, y_k는 $y_k=\sum_{i\in k} y_i$를 의미하며, k는 가구 총 수, i는 구성원을 나타낸다.

LFS에서 적용하고 있는 마지막 단계의 가중치로써 사후층화 가중치(g-factor)를 들 수 있다. 사후층화 가중치는 사후층화를 통한 보조정보로부터 획득하며 회귀추정방법을 이용하여 산출한다. 각 주 단위의 성별-연령대별 그룹, ER 지역과 CMA 지역에 대한 인구 총계, 센서스 결과에 의한 인구 추계 정보 등이 보조정보로 활용되었다. 추가적인 논의를 위해 다음과 같은 기호를 정의하기로 하자.

y_i : i번째 조사자에 대한 특성치,

y_k : k번째 가구에 대한 특성치 총계,

Q : 추정에 이용된 보조변수의 수, $q=1,2,\cdots,Q$,

x_{qi} : 조사자 i에 대한 q번째 지표변수의 값, 지표변수는 조사자 i가 j번째 범주에 속할 경우 1, 기타 0의 값을 갖는다.

x_{qk} : k번째 가구단위에 속하는 조사자들에 대한 q번째 지표변수값의 총계,

x_k : q번째 원소가 x_{qk}인 $Q\times 1$ 벡터,

c_k : k번째 가구의 크기,

$\hat{t}_{ya}$: 위에서 언급한 부차가중치에 근거한 추정치,

$\hat{t}_{x_qa}$: q번째 보조변수에 대한 부차가중치에 근거한 추정치.

사후층화 가중치를 산출하기 위해 $\hat{t}_{yr}=\hat{t}_{ya}+\sum_{q=1}^{Q}\hat{B}_q(t_{x_q}-\hat{t}_{x_qa})$와 같은 회귀추정량을 이용한다. 여기에서

$$\hat{t}_{x_qa}=\sum_s x_{qi}a_i\,,$$

$$\hat{B}=(\hat{B}_1,\cdots,\hat{B}_Q)^T=\Big(\sum_{k=1}^{n}\frac{x_kx_k^Ta_k}{c_k}\Big)^{-1}\sum_{k=1}^{n}\frac{x_ky_ka_k}{c_k}$$ 이며,

$(\sum_{k=1}^{n}\frac{x_k x_k^T a_k}{c_k})^{-1}$은 $Q\times Q$ 행렬, $\sum_{k=1}^{n}\frac{x_k y_k a_k}{c_k}$는 $Q\times 1$ 벡터를 나타낸다.

회귀추정량 $\hat{t}_{yr}$은 $\hat{t}_{yr}=\sum_{k\in s} y_k a_k g_k$와 같이 사후층화 가중치를 포함하는 식으로 재표현 할 수 있다. 여기서,

$$g_k=1+(t_x-\hat{t}_{xa})^T(\sum_{k\in s}\frac{x_k x_k^T a_k}{c_k})^{-1}\frac{x_k}{c_k}$$

로써 일명 g-factor로 불리우는 사후층화 가중치이다. 가구 구성원에 대한 사후층화 가중치는 $g_i=1+(t_x-\hat{t}_{xa})^T(\sum_{i\in s} z_i z_i^T a_i)^{-1}z_i$이며, 여기에서 $z_i=\frac{1}{c_k}\sum_{i=1}^{a_i}x_i$이다. 즉 사후층화 가중치의 특징은 가구에 대한 가중치와 가구 내의 구성원에 대한 가중치가 일치하여 모든 구성원들이 동일한 가중치를 갖는다는 점이다.

LFS의 최종가중치는 설계 가중치, 무응답 조정 가중치와 사후층화 가중치의 곱으로 표현되며, 이러한 최종가중치를 이용하여 주 지역 및 전국 단위의 경제활동인구 총계, 취업자 총계, 실업자 총계, 취업 및 실업률 등이 추정된다.

LFS에서 추정량의 분산계산은 잭나이프 방법을 이용한다. 좀 더 일반적인 경우의 잭나이프 방법에 대한 기술은 Wolter(1985)를 참조할 수 있으며, 여기에서는 LFS에서 적용하는 잭나이프 알고리즘을 소개하기로 한다.

(i) 잭나이프 방법을 적용하기 위해 h번째 층은 J_h개의 반복표본을 갖는다고 가정한다($a=1,2,\cdots,J_h$). 우선 특정 반복표본에 해당하는 모든 가구들을 제거한다. 여기에서 해당 표본에서 반복 총계는 $J=\sum_{h=1}^{H}J_h$이고, H는 해당 표본에서 층의 총계를 나타낸다.

(ii) 주어진 층에서 나머지 J_h-1개의 반복표본의 모든 가구들에 대해 부차가중치에 대한 보정이 이루어진다. 보정된 가중치의 값은

$a_k^{adj} = \frac{J_h}{(J_h-1)} a_k$ 이다.

(iii) 보정된 부차가중치와 남아있는 해당표본을 이용하여 관심 추정치 $\hat{t}_{yr(ha)}$ 를 계산하기 위한 최종가중치를 계산한다. 여기에서 (ha) 는 h 번째 층으로부터 a 번째 반복이 제거되었다는 것을 나타낸다.

해당표본의 모든 반복에 대해서 위의 (i)~(iii)의 절차가 반복되며, 결과로써 관심 추정치에 대한 J 개의 서로 다른 추정값을 얻게된다. 이러한 추정값을 이용하여 추정값의 분산을 계산하며 다음과 같은 분산 추정공식을 이용한다.

$$\hat{V}(\hat{t}_{yr}) = \sum_{h=1}^{H} \frac{(J_h-1)}{J_h} \sum_{a=1}^{J_h} (\hat{t}_{yr(ha)} - \hat{t}_{yr})^2 .$$

실업률에 대한 분산 추정은 다음의 추정공식을 이용할 수 있다.

$$\hat{V}(100\frac{\hat{t}_{yr}}{\hat{t}_{zr}}) = 100^2 \sum_{h=1}^{H} \frac{(J_h-1)}{J_h} \sum_{a=1}^{J_h} (\frac{\hat{t}_{yr(ha)}}{\hat{t}_{zr(ha)}} - \frac{\hat{t}_{yr}}{\hat{t}_{zr}})^2 ,$$

여기에서 y는 실업인구 총계, z는 경제활동인구 총계를 나타내며, 위의 결과는 실업률 $100(y/z)\%$ 에 대한 잭나이프 분산 추정공식이다.

월 변화량의 추정치에 대한 잭나이프 분산추정을 다음과 같이 산출할 수 있다. 연속되는 두 달의 월 추정치로부터 다음의 차분추정치 $\hat{D}_{yr} = \hat{t}_{yr}^{\,2} - \hat{t}_{yr}^{\,1}$ 를 고려하자. 여기에서 윗 첨자는 연속되는 월을 나타낸다. 대응되는 잭나이프 추정치는 $\hat{D}_{yr(ha)} = \hat{t}_{yr(ha)}^{\,2} - \hat{t}_{yr(ha)}^{\,1}$ 으로 표현할 수 있다. 이때 분산 추정치는 다음과 같이 주어질 수 있다.

$$\hat{V}(\hat{D}_{yr}) = \sum_{h=1}^{H} \frac{(J_h-1)}{J_h} \sum_{a=1}^{J_h} (\hat{D}_{yr(ha)} - \hat{D}_{yr})^2 .$$

LFS의 연속된 두 달의 조사에서는 5/6의 표본이 일치한다. 공통표본을 이용하여 두 달 간의 월 변화량의 차를 추정하는 것이 위의 추정방법에 비해 훨씬 효율적일 수 있다. Singh et al.(1997)은 이러한 공통표본을 이용하여 다

음과 같은 합성추정량을 제안하였다.

$$est^{C}_{(t+1)} = K \times est_{(t+1)} + (1-K) \times [est^{C}_{(t)} + change_{common}],$$

여기에서 윗첨자 C는 복합추정법을 의미하며, $change_{common}$은 공통표본을 이용한 변화량을 나타낸다. 이 추정량은 1998년 현재 캐나다 통계청에서 채택하고 있지는 않지만 새로운 표본설계에서는 사용될 전망이다.

위와 유사한 방법으로 n개의 월 추정치의 평균에 대한 분산추정을 고려할 수 있다. n개의 월 추정치의 평균은 $\hat{A}_{yr} = \sum_{i=1}^{n} \frac{\hat{t}_{yr}{}^{i}}{n}$ 이고, 이에 대응하는 잭나이프 추정값은 $\hat{A}_{yr(ha)} = \sum_{i=1}^{n} \frac{\hat{t}_{yr(ha)}{}^{i}}{n}$ 로 계산할 수 있으며, 추정치의 분산은 다음의 추정공식을 이용할 수 있다.

$$\hat{V}(\hat{A}_{yr}) = \sum_{h=1}^{H} \frac{(J_h - 1)}{J_h} \sum_{a=1}^{J_h} (\hat{A}_{yr(ha)} - \hat{A}_{yr})^2.$$

3.3.6 데이터 품질관리

모든 표본조사에서와 마찬가지로 LFS 추정값들도 표본 오차와 비표본 오차를 수반한다. 따라서 조사 추정값들이 올바르게 해석되기 위해서는 추정값들의 정도를 나타내는 측도에 관한 점검이 요구된다.

표본오차에 직접적으로 영향을 미치는 것은 표본크기라 할 수 있다. 일반적으로 표본크기가 증가함에 따라 표본오차는 감소한다. 표본의 크기뿐만 아니라 모집단의 변동, 추정 및 표본설계의 방법과 같은 요인들이 표본오차에 관련된 요인들이라 할 수 있다. 표본오차는 층화방법, 표본할당, 추출단위의 선택에서 뿐 만 아니라 다 단계 표본설계에서 매 단계에서 선택된 표본 추출방법 및 추정방법 등과 같은 요인들에 크게 의존한다.

표본설계 및 추정방법에 대한 효율성을 점검하기 위한 측도로는 평균제곱오차(MSE)가 이용된다. MSE는 추정값과 모집단의 실제값과의 편차제곱

합의 평균으로 보통 정의된다. 표본오차와 관련된 또 다른 중요한 측도로써 변동계수(CV)가 이용된다. Y가 어떤 특성치에 대한 추정치이고, d가 추정치의 표준오차라 할때, CV값은 $(d/Y)\times 100$으로 정의된다. 또한 추정치의 신뢰구간은 표준오차 d로부터 추론될 수 있다. 한편 시간에 따른 표본설계의 낙후성을 평가하기 위한 지표로써 설계효과(design effect)가 이용된다. 설계효과는 기존 표본설계로부터 조사된 추정값의 분산과 단순임의표본으로부터 계산된 추정값의 분산의 비로써 정의되며, LFS에서는 표본설계효과(sample design effect)와 전체설계효과(overall design effect)와 같은 두 가지 형태의 설계효과를 계산한다. 표본설계효과는 모총계에 대한 가중치의 조정없이 부차가중치만을 이용하여 계산되며, 전체설계효과는 앞서 언급되었던 최종가중치를 반영하여 계산된다. 따라서 표본설계효과는 표본설계의 효율성만이 반영되며, 전체설계효과는 층화, 다단계추출, 사후층화 및 추정 등의 표본설계의 전반적인 사항들이 반영된다고 보면 된다.

1997년 1월부터 7월까지의 LFS 조사자료에 대한 주 단위 및 전국 단위의 고용 및 실업관련 추정치들의 CV 값이 다음 <표3.12>에 주어졌다. 고용과 실업에 대한 월 변화 추정값의 표준오차는 <표3.13>에, 고용과 실업에 관한 설계효과는 <표3.14>에 주어졌다.

<표 3.12> LFS 고용 및 실업 추정치들의 CV 값 (1997)

Province	Employed CV(%)	Unemployed CV(%)	Province	Employed CV(%)	Unemployed CV(%)
Newfoundland	2.2	6.1	Manitoba	0.91	6.5
Prince Edward Island	1.7	6.5	Saskatchewan	1.1	7.4
Nova Scotia	1.2	5.3	Alberta	0.76	5.9
New Brunswick	1.2	5.5	British Columbia	0.90	5.1
Quebec	0.79	3.5	Canada	0.32	1.72
Ontario	0.54	3.0			

<표 3.13> LFS 고용 및 실업에 대한 월변화 추정값의 표준오차(1997)

Unit: thousands

Province	SE (employed)	SE (Unemployed)	Province	SE (employed)	SE (Unemployed)
Newfoundland	3	2	Manitoba	4	3
Prince Edward Island	1	1	Saskatchewan	3	2
Nova Scotia	4	3	Alberta	9	6
New Brunswick	3	2	British Columbia	12	9
Quebec	18	14	Canada	32	24
Ontario	20	15			

<표 3.14> LFS 고용 및 실업에 대한 설계효과 (1997)

Province	Employed		Unemployed	
	Sample	Overall	Sample	Overall
Newfoundland	2.7	0.83	1.4	1.3
Prince Edward Island	2.0	0.53	1.1	1.1
Nova Scotia	2.2	0.51	1.2	1.1
New Brunswick	2.0	0.56	1.4	1.4
Quebec	2.1	0.55	1.1	1.0
Ontario	3.3	0.50	1.2	1.1
Manitoba	2.2	0.41	1.1	1.1
Saskatchewan	2.4	0.63	1.2	1.2
Alberta	4.1	0.40	1.1	1.1
British Columbia	2.1	0.50	1.2	1.1
Canada	2.8	0.51	1.2	1.1

비표본오차는 표본조사의 매 단계에서 발생할 수 있으며 주로 조사원의 무관심, 오해 및 잘못된 해석 등의 사유에 기인하며, 추정값의 편향 및 변동에 직접적인 영향을 미친다. 관측값의 수가 많거나 혹은 대영역의 조사에서

는 비표본오차에 기인한 효과는 무시될 수도 있는 양이나, 소지역 추정의 문제에서는 민감한 문제로 인식되어진다. 비표본 편향 및 분산은 조사원에 대한 교육 및 조사원의 태도, 설문지 설계 상의 문제 또는 무응답을 처리하기 위해 이용되는 대체방법 등에서 발생될 수 있으며, 여기에서는 LFS에서의 적용범위 오차(coverage error), 무응답 오차, 빈집 오차(vacancy errror), 응답 오차(response error), 처리과정 오차(processing error)등에 관해서 설명하기로 한다.

적용범위 오차는 표본추출틀의 조사단위들이 목표모집단을 제대로 반영하지 못할 경우 발생할 수 있다. 조사단위들이 표본추출틀에서 누락되어 있는 경우, 목표모집단에 속하지 않은 단위들이 표본추출틀에 포함되어 있는 경우 또는 조사단위들이 표본추출틀에 중복되어 있는 경우 등이 적용범위 오차를 유발할 수 있는 일반적인 유형들이며, 이 중 조사단위들이 표본추출틀에서 누락되어 있는 경우가 LFS에서 가장 빈번히 발생하는 유형이라 할 수 있다. 나머지 유형의 문제는 LFS에서는 거의 무시된다. LFS에서는 적용범위 오차를 측정하기 위한 지표로써 손실률(slippage rate)을 이용한다. 손실률은 LFS 인구 추정치와 최근의 센서스 인구 추정치와의 차이에 대한 센서스 인구 추정치의 비율로써 정의된다. LFS에서는 CMA 지역, ER 지역, 주 및 전국 단위와 캐나다 지역의 성별(남, 녀)-연령대별(15-19, 20-24, 25-29, 30-39, 40-54, 55+) 범주에 대한 손실률을 매월 정기적으로 작성하고 있다. 손실률로부터 발생하는 비표본오차에 대한 보정은 추정과정에서 처리하고 있다. LFS 조사에서 평균적인 손실률의 양은 다음 <표3.15>와 같이 주어진다.

<표 3.15> LFS 평균 손실률(Average Slippage Rate(%))

Provinces		Average	Provinces	Average
Canada	all	9.3	Nova Scotia	8.6
	15-19세	6.1	New Brunswick	10.4
	20-24세	15.6	Quebec	8.0
	25-29세	16.1	Ontario	9.7
	30-39세	9.8	Manitoba	6.1
	40-54세	8.0	Saskatchewan	10.7
	55이상	6.9	Alberta	7.4
Newfoundland		9.8	British Coumbia	12.4
Prince Edward Island		11.6		

조사가구에 대한 무응답 발생요인으로써 가구 구성원의 부재, 가구 구성원의 비 정상적인 거주 환경, 인터뷰 거절 등의 요인을 들 수 있다. 여기에서 인터뷰 거절의 비율은 매월 조사에서 1~2% 정도로 매우 낮게 나타나며, 주 지역에 대해서도 월 조사와 비슷한 비율을 보이나 높게는 약3% 정도까지 나타난다. 단위 무응답에 대해서는 바로 전 달의 정보를 이용할 수 있다면 이를 이용하여 대체되며, 항목 무응답에 대해서는 표본 대체법이 이용된다. 가구 구성원이 거주하고 있지 않는 결측 가구 및 건물 철거 등으로 인한 비 존재 가구들에 대해서 발생하는 무응답은 편향에 영향을 미치지는 않지만 표본 분산에는 영향을 미치게 되므로 LFS에서는 이러한 유형의 오차 정보를 파악하기 위한 VC(vacancy check) 프로그램을 운영하고 있다. 다음 <표3.16>은 1997년 LFS 조사에서 발생한 평균 무응답률을 나타낸다.

<표 3.16> LFS 평균 무응답률(1997)

Provinces	Average(%)	Maximum(%)	Minimum(%)
Newfoundland	4.2	5.4	3.0
Prince Edward Island	3.5	4.8	2.4
Nova Scotia	6.3	7.3	4.6
New Brunswick	4.6	5.4	3.1
Quebec	5.4	6.6	3.7
Ontario	4.8	5.7	3.7
Manitoba	3.6	5.4	2.1
Saskatchewan	3.6	4.6	2.4
Alberta	4.9	6.3	3.1
British Columbia	5.7	6.7	4.5
Canada	4.9	5.5	3.8

LFS에서 결측가구(dwelling vacant)는 사람이 거주하고 있지 않는 가구, 계절 가구 또는 공사 중인 가구로 정의되어 분류된다. 철거 또는 조사 가구가 상점 등으로 용도가 변경된 경우는 비 존재 가구(dwelling non-existence)로 분류된다. 결측가구로 확인된 가구들은 LFS 추정치의 편향에 영향을 미치지는 않지만 표본 조사단위가 줄어들므로 추정 분산은 커지게 된다. 결측가구들은 새로운 입주자들이 상주할 가능성이 항상 존재하므로 매월 조사 대상에 포함된다. 그러나 LFS 조사에서 비 존재 가구로 확인된 조사가구들은 표본추출틀에서 일제히 삭제된다. 1997년 LFS 조사에서 발생한 평균적인 결측가구에 대한 비율이 다음 <표3.17>에 주어졌다.

<표 3.17> LFS 평균 결측률(vacant rate)(1997)

Provinces	Average(%)	Maximum(%)	Minimum(%)
Newfoundland	15.4	14.9	16.4
Prince Edward Island	20.5	18.6	23.0
Nova Scotia	16.8	15.2	18.7
New Brunswick	14.1	13.5	15.2
Quebec	14.0	11.9	15.8
Ontario	10.8	10.0	11.3
Manitoba	17.1	16.4	17.7
Saskatchewan	14.7	12.5	15.5
Alberta	8.7	8.1	9.8
British Columbia	9.5	8.7	9.8
Canada	13.0	12.2	13.5

응답 오차(response error)는 설문지 설계, 문항 구성, 응답자의 인지력, 인터뷰 방식, 조사가 수행되는 상황 및 조사정보가 수집되고 집계되는 과정 등에 기인할 수 있다. 조사정보가 수집되고 집계되는 과정에서 발생하는 응답오차는 CAI 시스템에 의해 어느 정도 보완되었다고 볼 수 있다.

처리과정 오차(processing error)는 자료 획득, 편집, 코딩, 가중치를 산출하는 과정 및 목록화 작업 등의 매 단계에서 발생할 수 있다. LFS 조사에서는 이러한 매 단계의 처리과정을 전산화 작업으로 통합하여 자료 처리과정에서 발생하는 오차를 최소화하고 있다. 전산화 통합 모드는 1993년부터 채택되어 시행되고 있으며 앞서 언급되었던 CAI 모드는 조사행정에서 조사원들의 조사과정을 보조하는 일종의 컴퓨터 보조관리 시스템을 말한다.

3.3.7 소지역 추정법

캐나다 노동력 조사에서는 표본설계 단계에서 EIER 지역과 CMA 지역과 같은 소지역 단위에 대한 추정을 고려하여 층화, 표본추출, 표본 배정 등

이 이루어진다. 소지역 통계 추정을 위한 추정량은 크게 설계 기반 추정량(design-based estimator), 간접추정량(indirect estimator), 모형 기반 추정량(model-based estimator)이 이용된다. 소지역 통계 작성 시 설계 기반 추정량이 목표 요구정도를 만족한다면 우선적으로 설계 기반 추정량을 이용하며 그렇지 못할 경우에는 추정량의 신뢰도를 확보할 수 있는 다른 추정방법을 이용한다.

(1) 설계 기반 추정량(Design-Based Estimator)

일반적으로 설계 기반 추정량은 직접추정량(direct estimator)과 수정된 직접추정량(modified direct estimator)으로 구분된다. 관심변수와 밀접한 관련이 있는 보조정보가 있는 경우에 이를 이용하는 사후층화추정량(post stratified estimator), 비추정량(ratio estimator), 회귀추정량(regression estimator) 등은 직접추정량의 일종이다. 직접추정량은 편향이 없는 추정량이지만 해당 소지역에 배정된 표본의 크기가 작은 경우에는 추정량의 분산이 커져서 신뢰성이 떨어지게 된다. 한편 수정된 직접추정량(modified direct estimator)은 해당 소지역 이외의 다른 지역의 조사결과를 추정과정에 추가적으로 이용하며 추정량의 불편성은 근사적으로 유지된다.

직접추정량(direct estimator)은 보통 해당 소지역에서 조사된 자료만을 이용하여 추정되며, 간혹 센서스나 행정자료로부터 획득된 보조정보를 조사자료에 추가하여 추정되기도 한다. 가장 간단한 총계추정에 대한 직접추정량으로써 다음과 같은 단순추정량(expansion estimator)을 들 수 있다.

$$\hat{Y}_{e,a} = \sum_{i \in s_a} \omega_i y_i \quad , \tag{3.25}$$

여기에서 s_a는 소지역 a의 표본들의 집합, ω_i는 조사단위 i에대한 가중치를 나타낸다. 위의 직접추정량은 불편추정량이나 소지역 a의 표본크기가 작을 경우에는 분산이 커지기 때문에 신뢰성에 문제가 있을 수 있다.

소지역 a의 모집단의 크기 N_a를 알고있을 경우에는 다음과 같은 사후층화추정량이 이용될 수 있다.

$$\hat{Y}_{pst,a} = N_a \frac{\sum_{i \in s_a} \omega_i y_i}{\sum_{i \in s_a} \omega_i}$$

$$= N_a \frac{\hat{Y}_{e,a}}{\hat{N}_{e,a}}$$

$$= N_a \bar{y}_{e,a} \qquad (3.26)$$

위의 사후층화추정량은 먼저 언급된 단순추정량보다는 안정적이나 보다 복잡한 조사에서는 비추정편향(ratio estimation bias)이 발생할 가능성이 있다.

표본이 층화추출되고 층 h에서 소지역 a의 모집단의 크기 $N_{h,a}$가 알려져 있을 경우에는 다음과 같은 유형의 사후층화추정량이 소지역 추정에 이용될 수 있다.

$$\hat{Y}_{st,pst,a} = \sum_h \left(N_{h,a} \frac{\sum_{i \in s_{h,a}} \omega_i y_i}{\sum_{i \in s_{h,a}} \omega_i} \right)$$

$$= \sum_h N_{h,a} \frac{\hat{Y}_{h,e,a}}{\hat{N}_{h,e,a}}$$

$$= \sum_h N_{h,a} \bar{y}_{h,a} \quad . \qquad (3.27)$$

여기에서 층 h는 표본설계 시 반영된 층이라기 보다는 사후층화에 의해 형성된 층을 말한다.

비추정법(ratio estimation)은 사후층화추정법과 유사하나 모집단 총계 N_a와 $N_{h,a}$대신에 보조정보에 의해 획득된 소지역 총계 X_a와 $X_{h,a}$를 이용하며, 이 값들을 알고 있을 경우 비추정량은 다음과 같이 정의된다.

$$\hat{Y}_{r,a} = X_a \hat{R}_a \ , \quad \hat{Y}_{st,r,a} = \sum_h X_{h,a} \hat{R}_{h,a} \ , \qquad (3.28)$$

여기에서 $\hat{R}_a = \hat{Y}_{e,a} / \hat{X}_{e,a}$는 Y_a / X_a의 추정값, $\hat{R}_{h,a} = \hat{Y}_{h,e,a} / \hat{X}_{h,e,a}$를

나타낸다.

회귀추정법(regression estimation)이 소지역 총계 추정에 이용되기도 한다. 이 방법은 관심변수 y와 공변량 x사이의 관계에서 회귀모수를 추정하여 소지역 총계 추정에 이용하는 방법으로써 추정량은 다음과 같은 형태로 주어진다.

$$\hat{Y}_{reg,a} = \hat{Y}_a + \hat{\beta}_a(X_a - \hat{X}_a) , \qquad (3.29)$$

여기에서 $\hat{Y}_a$는 직접추정 또는 사후층화추정법에 의해 추정된 소지역 a에 대한 총계 추정값이며 $\hat{X}_a$는 보조정보를 통해 $\hat{Y}_a$과 유사한 방법으로 추정된다. 추정모수 $\hat{\beta}_a$은 관심변수 y와 공변량 x의 관계로부터 추정되며 $\hat{\beta}_a = \sum_{i \in s_a} \nu_i^{-1} \omega_i y_i x_i^T (\sum_{i \in s_a} \nu_i^{-1} \omega_i x_i x_i^T)^{-1}$와 같이 주어진다. 여기에서 ν_i는 회귀가중치로써 주어지는 값이며, x가 상수이고 $\nu_i = x_i$일 경우에는 $\hat{\beta}_a = \hat{R}_a$인 관계가 성립한다. 회귀추정량의 불편성은 $\hat{Y}_a$와 $\hat{X}_a$의 불편성에 의존한다.

한편, 회귀추정량을 변형한 일종의 수정된 직접추정량(modified direct estimator)이 소지역 특성치 추정에 이용되기도 한다. 수정된 직접추정량은 해당 지역 외의 조사자료를 특성치 추정에 이용하며, 추정량의 불편성은 회귀추정량과 마찬가지로 근사적으로 만족된다. 예를 들면 식(7.5)에서 추정 회귀모수 $\hat{\beta}_a$ 대신에 회귀모수에 대한 합성추정량의 일종인 $\hat{\beta} = \sum_{i \in s} \nu_i^{-1} \omega_i y_i x_i^T (\sum_{i \in s} \nu_i^{-1} \omega_i x_i x_i^T)^{-1}$이 이용되었다면 이러한 추정량을 수정된 직접추정량(modified direct estimator)이라 부른다. 일반적으로 소지역 추정시 $\hat{\beta}$이 $\hat{\beta}_a$보다 안정적인 것으로 알려져 있으며, $\hat{\beta}$과 $\hat{\beta}_a$의 가중평균 $\lambda_a \hat{\beta}_a + (1-\lambda_a)\hat{\beta}$이 추정 회귀모수로 이용되기도 한다. 여기에서 λ_a는 적절히 선택되는 값이다. x가 상수이고 $\nu_i = x_i$인 경우에는 $\hat{\beta}$대신 $\hat{R} = \hat{Y}_e / \hat{X}_e$

이 이용될 수도 있다.

(2) 간접추정량(Indirect Estimator)

간접추정량은 합성추정량(synthetic estimator), 복합추정량(composite estimator), 표본수 의존 복합추정량(sample size dependent estimator) 등의 유형으로 구분되며, 해당 지역의 조사자료뿐만 아니라 해당 지역을 포함하고 있는 더 큰 지역의 조사자료를 소지역 추정과정에 이용하여 소지역 추정의 신뢰성을 확보하는 방법이다.

합성추정법(synthetic estimation)은 소지역 추정 시 소지역을 포함하는 대영역의 정보를 함께 이용하는 방법으로써 소지역과 대영역의 특성 구조가 유사하다는 가정 하에서 이용된다. 합성추정량의 분산은 직접추정량의 분산에 비해 작으나 전제한 가정이 성립하지 않을 경우에는 심각한 편향이 발생할 수 있다.

소지역의 특성치 평균이 전체 지역의 특성치 평균과 같다는 가정 하에서 만들어진 가장 간단한 형태의 합성추정량은 다음과 같다.

$$\hat{Y}_{syn,m,a} = N_a \frac{\sum_{i \in s} \omega_i y_i}{\sum_{i \in s} \omega_i} = N_a \bar{y} \qquad (3.30)$$

층화 또는 사후층화에 근거한 합성추정량은 보통 다음과 같은 형태로 주어진다.

$$\hat{Y}_{syn,st,m,a} = \sum_h N_{h,a} \frac{\sum_{i \in s_h} \omega_i y_i}{\sum_{i \in s_h} \omega_i} = \sum_h N_{h,a} \bar{y}_h \qquad (3.31)$$

직접추정법에서와 마찬가지로 합성추정법에서도 비합성추정법(ration synthetic estimation)이 고려될 수 있다. 비합성추정법은 모집단의 크기 N_a 또는 $N_{h,a}$ 외에 소지역 추정을 위한 보조정보로써 공변량 x를 이용하며 추정량(ratio synthetic estimator)은 다음과 같은 형태로 정의된다.

$$\hat{Y}_{syn,r,a}=X_a\frac{\hat{Y}_e}{\hat{X}_e}\ ,\quad \hat{Y}_{syn,st,r,a}=\sum_h X_{h,a}\frac{\hat{Y}_{h,e}}{\hat{X}_{h,e}}\ , \qquad (3.32)$$

여기에서 $\hat{Y}_e=\sum_{i\in s}\omega_i y_i$는 y에 대한 모집단 총계 추정량, $\hat{Y}_{h,e}=\sum_{i\in s_h}\omega_i y_i$는 층의 총계 추정치를 나타낸다. 비합성추정량들은 Gonzalez(1973), Gonzalez and Waksberg(1973), Ghangurde and Singh(1977, 1978)에 자세히 소개되어 있다. 한편, Singh and Tessier(1976)는 (3.32)식의 $\hat{Y}_{syn,r,a}$에서 $\hat{X}_e$ 대신에 X를 이용한 비합성추정량의 대체식 $\hat{Y}_{syn,r,a}=X_a\,\hat{Y}_e/X$을 제안하였다. 여기에서 $\hat{Y}_{syn,r,a}$와 $\hat{Y}_{syn,r,a}$는 모두 같은 양의 편향을 가지며, $\hat{Y}_{syn,r,a}$의 편향은 표본의 크기가 클 경우에는 무시될 수 있다. 두 추정량 중 하나의 추정량을 선택하는 문제에서는 $\hat{Y}_e$와 $\hat{X}_e$의 상관계수 ρ를 참조하도록 하였다. 일반적으로 표본의 크기가 클 경우, 두 추정량의 분산은 상관계수의 값이 $\rho\geq 0.5c_x/c_y$일때 $V(\hat{Y}_{syn,r,a})\leq V(\hat{Y}_{syn,r,a})$인 관계가 성립한다. 여기에서 c_x와 c_y는 각각 $\hat{X}_e$와 $\hat{Y}_e$의 변동계수(coefficient of variation)를 나타낸다. 상관계수 ρ의 값이 크거나 모집단의 분포가 한쪽으로 치우쳐져 있을 경우에는 $\hat{Y}_{syn,r,a}$가 선호되며, 변동계수 c_x의 값이 크거나 상관계수 ρ의 값이 적당할 경우에는 보통 $\hat{Y}_{syn,r,a}$를 선택한다.

소지역의 보조정보로써 이용된 공변량 x외에 추가적인 보조변수 z를 도입하여 소지역 특성치를 추정하는 다음과 같은 이변량 비합성추정량이 소지역 추정에 이용될 수 있다.

$$\hat{Y}_{syn,r,a}^{(2)}=\gamma_a X_a\frac{\hat{Y}_e}{\hat{X}_e}+(1-\gamma_a)Z_a\frac{\hat{Y}_e}{\hat{Z}_e}\ , \qquad (3.33)$$

여기에서 γ_a는 적절히 선택되는 값이다. 보다 일반적인 다변량 비합성추정량은 Olkin(1958)에서 참조할 수 있다.

회귀합성추정법은 비합성추정법과 유사하며 추정량은 다음과 같이 주어

진다.

$$\hat{Y}_{syn.reg.a} = \hat{\beta} X_a \ , \quad \hat{\beta} = \sum_{i \in s} \nu_i^{-1} \omega_i y_i x_i^T \left(\sum_{i \in s} \nu_i^{-1} \omega_i x_i x_i^T \right)^{-1} \qquad (3.34)$$

회귀합성추정법은 표본설계의 층 내에서 또는 사후층화에 의해 형성된 층 내에서도 응용이 가능하며, Royall(1979)은 이러한 내용을 반영한 다음과 같은 회귀합성추정량을 제안하였다.

$$\hat{Y}_{syn.Roy.a} = \sum_{i \in s_a} y_i + \hat{\beta} (X_a - \sum_{i \in s_a} x_i), \qquad (3.35)$$

복합추정량(composite estimator)은 직접추정량의 불안정성과 합성추정량의 잠재적 편향 가능성을 보완하기 위해 두 추정량의 가중평균을 취하며 일반적인 형태는 다음과 같이 주어진다.

$$\hat{Y}_{com.a} = \lambda_a \hat{Y}_{dir.a} + (1 - \lambda_a) \hat{Y}_{syn.a} \ , \qquad (3.36)$$

여기에서 가중치 λ_a는 적절히 선택되는 값이다.

가중치 λ_a는 결정하는 방법은 크게 세가지 정도로 구분될 수 있다. 첫 번째 방법은 가장 간단한 방법으로써 가중치 λ_a를 고정계수로 두는 방법인데 추정량의 신뢰성에 문제가 있어 많이 사용되지는 않는다. 두 번째 방법은 추정하고자하는 소지역의 표본크기를 반영하는 방법이다. 이 경우 가중치 λ_a는 $\hat{N}_{e.a}/N_a$의 함수로 표현된다. Drew et al.(1982)은 표본크기에 의존하는 복합추정량으로써 다음과 같은 추정량을 제안하였다.

$$\hat{Y}_{ssd.r.a} = \lambda_a \hat{Y}_{r.a} + (1 - \lambda_a) \hat{Y}_{syn.r.a} \ , \qquad (3.37)$$

여기에서 $\lambda_a = \begin{cases} 1 & , \text{ if } \hat{N}_{e.a} \geq \delta N_a \\ \hat{N}_{e.a} / \delta N_a & , \textit{otherwise} \end{cases}$ 이며, δ는 합성추정량 부분의 편향을 보정하기 위해 주관적으로 결정되는 값이다. 캐나다 노동력조사에서는 $\delta = 2/3$를 이용한다. 식 (3.37)의 복합추정량은 직접추정량의 신뢰도가 완전히 확보된 지역에 대해서는 합성추정량의 가중치가 0이 되기 때문에 이러한 지역의 경우 직접추정값이 곧 복합추정값으로 선택된다고 볼 수 있다. 그

렇지 않은 기타 지역에 대해서는 직접추정값과 합성추정값의 가중평균값으로 복합추정값이 계산된다. 캐나다 노동력조사에서 이러한 기타 지역들에 대한 합성추정량의 평균 가중치는 약 10% 정도이며 많아야 20% 를 초과하지는 않는다. 이때 δ의 값은 [2/3 , 3/2] 의 범위에 있는 것으로 알려져 있다. 이외의 표본크기 의존 복합추정량으로써 Sandal(1984)의 추정량 $\hat{Y}_{ssd,reg,a}=\lambda_a \hat{Y}_{sreg,a}+(1-\lambda_a)\hat{Y}_{syn,reg,a}$ 을 들 수 있다. 사용된 가중치는 $\lambda_a=\hat{N}_{e,a}/N_a$ 이다. Rao(1986)는 위와 동일한 추정량에 대해 가중치를 약간 달리 적용할 것을 제안하였다. Rao의 가중치는 $\hat{N}_{e,a}\geq N_a$ 인 지역에 대해서는 $\lambda_a=1$, 기타 지역에 대해서는 Sandal의 가중치와 동일하다. Sandal and Hidiroglou(1989)는 Rao의 가중치에서 $\hat{N}_{e,a}<N_a$ 일때 $\lambda_a=(\hat{N}_{e,a}/N_a)^{h-1}$ 를 사용할 것을 제안하였다. 여기에서 h는 합성추정량 편향을 감안하여 적절히 선택되는 값이다. 가중치를 결정하는 세 번째 방법은 직접추정량과 합성추정량의 평균제곱오차와 두 추정량의 공분산을 자료로부터 추정하여 최적가중치를 산정하는 방법이다. 복합추정량의 평균제곱오차는 다음 식과 같이 나타낼 수 있다.

$$MSE(\hat{Y}_{com,a})=\lambda_a^2 MSE(\hat{Y}_{dir,a})+(1-\lambda_a)^2 MSE(\hat{Y}_{syn,a}) +2\lambda_a(1-\lambda_a)E(\hat{Y}_{dir,a}-Y_a)(\hat{Y}_{syn,a}-Y_a) \quad (3.38)$$

$\hat{Y}_{com,a}$ 의 MSE를 최소화하는 가중치 λ_a 다음 식과 같이 주어질 수 있다.

$$\lambda_a=\frac{\widehat{MSE}(\hat{Y}_{syn,a})-E(\hat{Y}_{syn,a}-Y_a)(\hat{Y}_{dir,a}-Y_a)}{\widehat{MSE}(\hat{Y}_{syn,a})+\widehat{MSE}(\hat{Y}_{dir,a})-2E(\hat{Y}_{syn,a}-Y_a)(\hat{Y}_{dir,a}-Y_a)} \quad (3.39)$$

식 (3.39)에서 $\hat{Y}_{dir,a}$ 와 $\hat{Y}_{syn,a}$ 의 공분산의 항이 $\widehat{MSE}(\hat{Y}_{syn,a})$ 와 $\widehat{MSE}(\hat{Y}_{dir,a})$ 에 비해 매우 작다고 가정할 수 있다면 다음과 같은 근사적인 가중치를 이용할 수도 있다.

$$\lambda_a^* = \frac{\widehat{MSE}(\hat{Y}_{syn,a})}{\widehat{MSE}(\hat{Y}_{syn,a}) + \widehat{MSE}(\hat{Y}_{dir,a})} \qquad (3.40)$$

(3) 모형 기반 추정량(Model-Based Estimator)

소지역 추정에 자주 이용되는 모형 기반 추정법(model-based estimation)으로는 EBLUP (empirical best linear unbiased prediction), EB(empirical Bayes), HB(hierarchical Bayes) 접근법 등이 있다. 최근에는 소지역 통계 작성을 위해 횡단면 조사자료(cross-sectional data)와 시계열자료(time series data)를 함께 추정과정에 이용하는 방법에 관한 연구가 활발히 진행되고 있다. 이 절에서는 횡단면 조사자료를 이용한 모형기반 추정량과 횡단면 조사자료와 시계열자료를 함께 이용하는 모형기반 추정량에 대한 최근의 연구들을 소개하기로 한다.

y_i를 i번째 소지역의 관심모수 θ_i에 대한 직접추정량, x_i를 모수 θ_i의 추정에 필요한 설명변수이고, 모형 $y_i = \theta_i + e_i$, $E(e_i) = 0$ 를 가정할때, 소지역 i에 대한 다음과 같은 선형회귀모형(linear regression model)을 고려할 수 있다.

$$\theta_i = \beta_0 + \beta_1 x_i, \qquad i = 1, 2, \cdots, I , \qquad (3.41)$$

여기에서 β_0와 β_1은 회귀모수를 나타낸다. 이때 θ_i에 대한 회귀합성추정량(regression synthetic estimator)은 다음과 같이 주어질 수 있다.

$$\hat{\theta}_{i(reg)} = \hat{\beta}_0 + \hat{\beta}_1 x_i , \quad i = 1, 2, \cdots, I , \qquad (3.42)$$

여기에서 $\hat{\beta}_0$와 $\hat{\beta}_1$은 결합모형 $y_i = \beta_0 + \beta_1 x_i + e_i$ ($i = 1, 2, \cdots, I$) 로부터 계산되는 최소제곱추정량을 나타낸다. 조사 추정량 y_i들의 공분산을 추정할 수 있을 경우에는 일반화 가중최소제곱추정량을 이용할 수도 있다. 위의 회귀합성추정량은 조사 추정량 y_i들에 대한 가중치가 반영되지 않기 때문에 큰 편향이 발생할 수 있다. 반면, EB(Empirical Bayes) 추정량이나

EBLUP(Empirical Bayes Linear Unbiased Predictor)는 적당한 가중치가 부여되어 편향 발생이 다소 억제되는 결과를 얻을 수 있다.

이러한 편향에 대한 문제점을 해결하기 위해 Fay and Herriot(1979)는 모형 (3.41)을 다음과 같이 해당 소지역에 대한 랜덤효과 v_i를 갖는 모형으로 보완하였다.

$$\theta_i = \beta_0 + \beta_1 x_i + v_i \ , \ i=1,2,\cdots,I \ , \tag{3.43}$$

여기에서 v_i는 평균이 0이고 분산이 σ_v^2을 갖는 서로 독립인 정규분포를 따르는 확률변수, e_i는 평균이 0이고 분산이 σ_i^2인 서로 독립인 정규 확률변수를 나타내며, σ_i^2은 기지인 값으로 가정된다. 이때 결합모형은 다음과 주어진다.

$$y_i = \beta_0 + \beta_1 x_i + v_i + e_i \ , \ i=1,2,\cdots,I \ . \tag{3.44}$$

위의 모형으로부터 θ_i의 EB 추정량은 직접조사추정량 y_i와 회귀합성추정량 $\hat{\theta}_{i(reg)} = \hat{\beta}_0 + \hat{\beta}_1 x_i$의 가중합으로 표현되며 다음 식과 같이 주어진다.

$$t_i(\hat{\sigma}_v^2, \boldsymbol{y}) = \omega_i y_i + (1-\omega_i)\hat{\theta}_{i(reg)} \ , \tag{3.45}$$

여기에서 $\omega_i = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \sigma_i^2)$, $\hat{\beta}_0$와 $\hat{\beta}_1$은 결합모형 하에서 추정된 가중최소제곱추정량을 나타내며, $\hat{\sigma}_v^2$은 σ_v^2의 적률추정량 또는 최대우도추정량 등이 이용될 수 있다. Fay and Herriot(1979)는 1970년 미국의 인구주택 총조사 자료로부터 인구 1000 미만의 소지역에 대한 소득관련 추정에 식 (3.45)의 EB 추정량을 이용하였고, EB 추정량이 직접 조사 추정량이나 합성추정량에 비해 표본오차가 작다는 사실을 수치적으로 제시하였다.

횡단면자료를 이용한 소지역 추정방법은 조사 시기가 상이한 조사자료들의 정보를 모형에 반영시키는 것은 사실상 어렵다. Scott et al.(1977), Jones(1980), Tiller(1989) 등은 이러한 단점을 보완하기 위해 반복적인 월별 조사자료들의 정보와 센서스 및 행정자료를 모형에 포함시킨 횡단면 시계열

모형들을 소지역 추정 문제에 도입하였다.

θ_{it}, y_{it}와 x_{it}를 각각 조사시기 t에서 소지역 i에 대한 모평균, 직접 조사추정값, i번째 소지역과 관계가 있는 연관변수라 할때, 우선 다음과 같은 모형을 고려한다.

$$y_{it}=\theta_{it}+e_{it}\ ,\quad i=1,2,\cdots,I\ ,\quad t=1,2,\cdots,T\ , \tag{3.46}$$

여기에서 표본오차 e_{it}의 평균은 0, 분산공분산행렬은 기지인 블럭대각행렬 Σ_i ($T\times T$ 행렬)로 가정한다. 모평균 θ_{it}에 관한 모형은 다양한 유형으로 설정될 수 있으며 다음과 같은 모형들이 고려될 수 있다.

(Ⅰ) $\theta_{it}=\beta_0+\beta_1 x_{it}+v_i+\varepsilon_{it}\ ,\quad i=1,2,\cdots,I\ ,\quad t=1,2,\cdots,T\ ,$

여기에서 v_i는 소지역 고정효과, $\varepsilon_{it} \overset{\text{ind}}{\sim} N(0\ ,\ \sigma^2)$.

(Ⅱ) $\theta_{it}=\beta_0+\beta_1 x_{it}+v_i+\varepsilon_{it}\ ,\quad i=1,2,\cdots,I\ ,\quad t=1,2,\cdots,T\ ,$

여기에서 $v_i \overset{\text{ind}}{\sim} N(0\ ,\ \sigma_v^2)$, $\varepsilon_{it} \overset{\text{ind}}{\sim} N(0\ ,\ \sigma^2)$, $\{v_i\}$와 $\{\varepsilon_{it}\}$는 서로 독립이며, 모형(Ⅰ)과는 달리 v_i들이 랜덤효과로 가정되었다.

(Ⅲ) $\theta_{it}=\beta_0+\beta_1 x_{it}+v_i+u_t+\varepsilon_{it}\ ,\quad i=1,2,\cdots,I\ ,\quad t=1,2,\cdots,T\ ,$

여기에서 $v_i \overset{\text{ind}}{\sim} N(0\ ,\ \sigma_v^2)$, $u_t \overset{\text{ind}}{\sim} N(0\ ,\ \sigma_u^2)$, $\varepsilon_{it} \overset{\text{ind}}{\sim} N(0\ ,\ \sigma^2)$, $\{v_i\}$, $\{u_t\}$와 $\{\varepsilon_{it}\}$는 서로 독립이다. v_i는 소지역에 대한 랜덤효과, u_t는 조사시기에 대한 랜덤효과이다.

(Ⅳ) $\theta_{it}=\beta_0+\beta_1 x_{it}+v_i+u_{it}\ ,\quad u_{it}=\rho u_{i,t-1}+\varepsilon_{it}\ ,\quad |\rho|<1\ ,$

여기에서 $v_i \overset{\text{ind}}{\sim} N(0\ ,\ \sigma_v^2)$, $\varepsilon_{it} \overset{\text{ind}}{\sim} N(0\ ,\ \sigma^2)$, $\{v_i\}$와 $\{\varepsilon_{it}\}$는 서로 독립이며, $\{u_{it}\}$는 $AR(1)$ 과정을 따른다. 모형 (Ⅳ)는 다음과 같이 시차모형(lag

model)으로 재표현 가능하다.

$$\theta_{it} = \rho\theta_{i,t-1} + (1-\rho)\beta_0 + \beta_1 x_{it} - \beta_1\rho x_{i,t-1} + (1-\rho)v_i + \varepsilon_{it} . \qquad (3.47)$$

모형 (Ⅳ)는 모평균 θ_{it}와 보조변수 x_{it}가 이전 조사시기에서의 값들을 모형에 반영시키기 때문에 위의 네가지 모형들 중에서는 가장 현실적인 추정 모형이라 할 수 있다. 따라서 모형 (Ⅳ)를 이용한 결합모형을 고려한다면 다음과 같이 주어진다.

$$y_{it} = \beta_0 + \beta_1 x_{it} + v_i + (e_{it} + u_{it}) , \qquad (3.48)$$

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it} , \quad |\rho| < 1 ,$$

여기에서 $v_i \overset{ind}{\sim} N(0, \sigma_v^2)$, $\varepsilon_{it} \overset{ind}{\sim} N(0, \sigma^2)$, e_{it}는 평균이 0이고, 기지인 블록 대각 공분산행렬 $\Sigma = diag(\Sigma_1, \Sigma_2, \cdots, \Sigma_I)$를 갖는다. 캐나다 노동력조사에서는 표본 공분산행렬 Σ를 이용할 수 없기 때문에 Tiller(1989)는 복합오차 $w_{it} = e_{it} + u_{it}$를 $AR(1)$ 과정으로 취급한 후 다음과 같은 결합모형을 이용하여 소지역 추정을 시도하였다.

$$y_{it} = \beta_0 + \beta_1 x_{it} + v_i + w_{it} , \qquad (3.49)$$

$$w_{it} = \rho w_{i,t-1} + \varepsilon_{it} , \quad |\rho| < 1 ,$$

여기에서 $\theta_{it} = \beta_0 + \beta_1 x_{it} + v_i$, $v_i \overset{ind}{\sim} N(0, \sigma_v^2)$, $\varepsilon_{it} \overset{ind}{\sim} N(0, \sigma^2)$.

$\{y_{it}\}$를 $\boldsymbol{y} = (y_{11}, \cdots, y_{1T}; y_{21}, \cdots, y_{IT})^T = (\boldsymbol{y}_1^T, \cdots, \boldsymbol{y}_I^T)^T$로 표현하면 위의 모형 (3.49)는 다음과 같은 일반적인 혼합모형(mixed model)의 일종으로 볼 수 있다.

$$\boldsymbol{y} = X\beta + Zv + w , \quad v \sim (0, \sigma_v^2 I), \quad w \sim (0, \sigma^2(I \otimes \Gamma)), \qquad (3.50)$$

여기에서 $X^T = (X_1^T, \cdots, X_I^T)$, $Z = I \otimes 1_T$, $\beta = (\beta_0, \beta_1)^T$, X_i는 t번째 행이 $(1, x_{it})$로 주어지는 $T \times 2$ 행렬, I는 크기 I인 항등행렬, 1_T는 1을

원소로 갖는 t-벡터, Γ는 (i, j)번째 원소가 $\gamma_{ij}=(1-\rho^2)^{-1}\rho^{|i-j|}$인 $T\times T$ 행렬을 나타낸다.

β와 v의 일차결합 $\tau=k^T\beta+m^Tv$에 대한 $\tilde{\tau}$ $(=\tilde{\theta}_{it})$의 BLUP(best linear unbiased predictor)와 BLUP의 평균제곱오차(MSE)는 다음과 같이 주어진다(Henderson, 1975).

$$\tilde{\tau}=k^T\hat{\beta}+m^TZ^T\Sigma^{-1}(y-X\hat{\beta})(\sigma_v^2/\sigma^2)\ , \tag{3.51}$$

$$MSE(\tilde{\theta}_{it})=\sigma^2\{k^T(X^T\Sigma^{-1}X)^{-1}k+(\sigma_v^2/\sigma^2)m^Tm-(\sigma_v^2/\sigma^2)^2m^TZ^T\Sigma^{-1}AZm$$
$$-2(\sigma_v^2/\sigma^2)k^T(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Zm\}\ , \tag{3.52}$$

여기에서 $\Sigma=I\otimes[(\sigma_v^2/\sigma^2)J+\Gamma]$, $\hat{\beta}=(X^T\Sigma^{-1}X)^{-1}(X^T\Sigma^{-1}y)$, J는 원소들이 1로 구성된 $T\times T$ 행렬, $A=I-X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$을 나타낸다.

식 (3.51)을 이용하여 결합모형 (3.49)의 $\theta_{it}(=\tau)$에 대한 BLUP 추정량을 계산하면 주요 항들은 다음과 같이 주어질 수 있다.

$$k^T=(1, x_{it}),\quad m^T=(0, \cdots, 0, 1, 0, \cdots, 0)_i\ ,$$

$$m^TZ^T\Sigma^{-1}(y-X\hat{\beta})=\ 1_T'\ [(\sigma_v^2/\sigma^2)J+\Gamma]^{-1}(y_i-X_i\hat{\beta})\ .$$

식 (3.51)의 BLUP는 미지인 분산비(variance ratio) σ_v^2/σ^2와 자기상관계수(autocorrelation) ρ에 의존하므로 이들 값들은 추정되어야 한다. 모수들에 대한 BLUP는 다음과 같이 주어진다(Pantula and Pollack, 1985).

$$\hat{\rho}=\Big[\sum_{i=1}^{I}\sum_{t=1}^{T-2}\hat{e}_{it}(\hat{e}_{i,t+1}-\hat{e}_{i,t+2}\Big]\Big[\sum_{i=1}^{I}\sum_{t=1}^{T-2}\hat{e}_{it}(\hat{e}_{it}-\hat{e}_{i,t+1}\Big]^{-1} \tag{3.53}$$

$\hat{\sigma}_v^2$과 $\hat{\sigma}^2$은 다음 (3.54), (3.55), (3.56), (3.57)식의 정의로부터 유도될 수 있다.

$$z_{it}^{(1)}=z_{it}-z_{it}^{(2)}\ , \tag{3.54}$$

여기서 $z_{it}=\begin{cases} y_{it}-\hat{\rho}y_{i,t-1}\ , & t\geq 2 \\ f_1y_{it}\ , & t=1 \end{cases}$, $z_{it}^{(2)}=c^{-1}d_if_t$, $c=(1-\hat{\rho})[T-(T-2)\hat{\rho}]$,

$f_t=\begin{cases}1-\hat{\rho}^2, & t=1\\ 1-\hat{\rho}, & t\geq 2\end{cases}$, $d_i=\sum_{t=1}^{T} f_t z_{it}$.

$$h_{0it}^{(1)}=h_{0it}-h_{0it}^{(2)}, \tag{3.55}$$

여기에서 $h_{0it}=\begin{cases}1-\hat{\rho}, & t\geq 2\\ f_1, & t=1\end{cases}$, $h_{0it}^{(2)}=c^{-1}d_i f_t$, $f_t=\begin{cases}1-\hat{\rho}^2, & t=1\\ 1-\hat{\rho}, & t\geq 2\end{cases}$,

$d_i=\sum_{t=1}^{T} f_t h_{0it}$.

$$h_{1it}^{(1)}=h_{1it}-h_{1it}^{(2)}, \tag{3.56}$$

여기에서 $h_{1it}=\begin{cases}x_{it}-\hat{\rho}x_{i,t-1}, & t\geq 2\\ f_1 x_{it}, & t=1\end{cases}$, $h_{1it}^{(2)}=c^{-1}d_i f_t$, $f_t=\begin{cases}1-\hat{\rho}^2, & t=1\\ 1-\hat{\rho}, & t\geq 2\end{cases}$,

$d_i=\sum_{t=1}^{T} f_t h_{1it}$.

$$g_i=\sum_{t=1}^{T} f_t z_{it},\ f_{0i}=\sum_{t=1}^{T} f_t h_{0it},\ f_{1i}=\sum_{t=1}^{T} f_t h_{1it}. \tag{3.57}$$

종속변수를 $z_{it}^{(1)}$, 설명변수를 $h_{0it}^{(1)}$와 $h_{1it}^{(1)}$를 갖는 절편항이 없는 회귀식을 적합시켰을 때 얻어지는 잔차제곱합을 $\hat{e}^T\hat{e}$라 하고, 종속변수 g_i, 설명변수 f_{0i}, f_{1i}를 갖는 절편항이 없는 회귀식에서 얻어지는 잔차제곱합을 $\hat{u}^T\hat{u}$라 할때, σ_v^2과 σ^2의 BLUP는 최종적으로 다음과 같이 주어진다.

$$\hat{\sigma}_v^2=c^{-1}(I-2)^{-1}[\,\hat{u}^T\hat{u}-\hat{\sigma}^2(I-2)\,],$$

$$\hat{\sigma}^2=[I(T-1)-2]^{-1}\hat{e}^T\hat{e}. \tag{3.58}$$

따라서 θ_{it}의 EBLUP $\hat{\theta}_{it}$는 식 (3.51)에서 $\hat{\rho}$, σ_v^2, σ^2을 대체하여 얻을 수 있다.

만약 $p-1(\geq 2)$개의 x변수들이 모형에 포함되어 있다면, θ_{it}의 EBLUP $\hat{\theta}_{it}$는 다음과 같이 추정하면 된다. 우선 y_{it}와 $x_{1it},\cdots,x_{p-1,it}$의 회귀식으로부터 $\{\hat{e}_{it}\}$을 추정한다.

다음으로 $\{h_{jit}, h_{jit}^{(1)}, h_{jit}^{(2)}, j=0,1,\cdots,p-1\}$을 앞서 언급된 바와 같이 1,

$x_{1it}, \cdots, x_{p-1,it}$의 원소들로 정의한 후, $z_{it}^{(1)}$과 $h_{0it}^{(1)}, h_{1it}^{(1)}, \cdots, h_{p-1,it}^{(1)}$의 절편항이 없는 회귀식으로부터 $\hat{e}^T\hat{e}$을 추정한다. 같은 방법으로 $f_{ji}=\sum_{t=1}^{T} f_t h_{jit}$ $(j=0,1,\cdots,p-1)$를 계산한 후, g_i와 $f_{0i}, f_{1i}, \cdots, f_{p-1,i}$의 절편항이 없는 회귀식을 적합시켜 $\hat{u}^T\hat{u}$를 구한다. 마지막으로 (3.58)식에서 $I(T-1)-2$를 $I(T-1)-p$로, $I-2$를 $I-p$로 대체하여 BLUP $\hat{\sigma}_v^2$, $\hat{\sigma}^2$과 $\hat{\rho}$를 계산하면 EBLUP $\hat{\theta}_{it}$을 얻을 수 있고, (3.52)식으로부터 EBLUP $\hat{\theta}_{it}$의 MSE 값을 계산할 수 있다. 한편, 모형 (3.49) 하에서 조사 추정량 y_{it}의 MSE는 다음과 같이 주어진다.

$$MSE(y_{it}) = E(y_{it}-\theta_{it})^2 = V(w_{it}) = \frac{\sigma^2}{(1-\rho^2)}. \qquad (3.59)$$

$MSE(y_{it})$의 추정량은 위의 (3.59)식에서 σ^2, ρ를 각각 $\hat{\sigma}^2$, $\hat{\rho}$로 대체하여 얻을 수 있다.

(4) 소지역 추정량들의 효율

$\hat{Y}_{M,a}(r)$을 소지역 추정법 M을 이용하여 추정된 r번째 반복에서 특성치 Y_a의 몬테카를로 추정값이라 하자. 이 때 n개의 소지역에 대한 평균제곱오차 추정값의 평균은 다음 식을 이용하여 계산할 수 있다.

$$Avg\ \widehat{MSE}_M = \frac{1}{n}\sum_a \sum_{r=1}^{R} \frac{(\hat{Y}_{M,a}(r)-Y_a)^2}{R}. \qquad (3.60)$$

소지역 추정법 M을 이용하여 추정된 추정량들의 효율을 직접추정법 M_0를 이용한 추정량과 비교하여 나타낸다면 상대 효율은 다음 식을 이용하여 구할 수 있다.

$$Eff(M\ vs\ M_0) = \frac{Avg\ \widehat{MSE}_M}{Avg\ \widehat{MSE}_{M_0}}. \qquad (3.61)$$

여기에서 $Avg\ \widehat{MSE}_{M_0}$는 직접추정법 M_0에 의해 추정된 추정값들에 대한 평균제곱오차의 평균을 나타낸다.

(5) LFS의 소지역 통계 작성방법

현재 캐나다 LFS에서 소지역 통계 작성을 위해 사용되고 있는 추정량은 Drew et al.(1982)에 기초한 표본수 의존 추정량(sample size dependent estimator)이다. 우선 소지역 a에 대해서 일반화 회귀추정량으로 관심변수의 총계를 추정한다.

$$\hat{Y}_{GREG,a} = \hat{Y}_{e,a} + \hat{\beta}_a(X_a - \hat{X}_{e,a}) , \qquad (3.62)$$

여기에서 $\hat{Y}_{e,a} = \sum_{i \in s_a} \omega_i y_i$, $\hat{X}_{e,a} = \sum_{i \in s_a} \omega_i x_i$, $\hat{\beta}_a = \sum_{i \in s_a} \omega_i y_i x_i^T (\sum_{i \in s_a} \omega_i x_i x_i^T)^{-1}$이다.

관심변수의 총계를 추정할 수 있으면 LFS에서 관심의 대상인 경제활동인구수, 실업자 수, 취업자 수, 실업률, 취업률 등을 모두 추정할 수 있게 된다. 추정과정에 이용된 가중치 ω_i는 설계 가중치와 무응답 가중치 조정이 반영된 것이며, LFS의 최종 가중치를 의미하는 것은 아니다. 일반적으로 ER 지역의 소지역 통계 작성에 사용되는 보조정보는 주(Province) 단위의 노동력 통계 작성에 이용될 수 있는 보조정보에 비해 제한적이라 할 수 있다.

Drew et al.(1982)는 LFS에 알맞은 소지역 추정량을 찾기 위해 여러 종류의 추정량들을 비교하였다. 여기에서 사용된 소지역은 CD(Census Division) 지역이고 각 소지역에 대해 다음과 같은 세 개의 범주에 대한 보조정보를 추정과정에 이용하였다.

(i) 연령 15~16세, 65세 이상 전체인구

(ii) 연령 17 ~64세 여성인구

(iii) 나이 17~64세 남성인구 .

소지역 a를 포함하는 주(Province) 단위에서 $\hat{\beta}$을 구하여 합성추정량 $\hat{Y}_{SYN.GREG.a}=\hat{\beta}X_a$을 계산하였다. 여기에서 $\hat{\beta}=\sum_{i\in s}\omega_i y_i x_i^T(\sum_{i\in s}\omega_i x_i x_i^T)^{-1}$이며, 주 단위에서 사용된 보조정보는 30개의 연령 및 성별그룹의 총 수, 각 ER 지역과 CMA 지역에 대한 인구 총계 등이다.

LFS에서 사용되고 있는 표본수 의존 추정량은 앞서 언급한 두 추정량의 가중평균 $\hat{Y}_{SSD.a}=\lambda_a\hat{Y}_{GREG.a}+(1-\lambda_a)\hat{Y}_{SYN.GREG.a}$ 을 취한다. 여기에서 가중치 λ_a는 $\lambda_a=\begin{cases}1, & \hat{N}_{e.a}\geq\delta N_a\\ \hat{N}_{e.a}/\delta N_a, & otherwise\end{cases}$ 이고, δ값은 2/3를 사용하고 있다. 만약 소지역에 대한 직접추정량이 목표 요구정도를 만족하고 있다면 합성추정량 부분에 대한 가중치는 0이 된다.

현재 LFS에서 소지역의 취업률과 실업률에 대한 추정값은 표본수 의존 추정량으로 구해진 추정치의 3개월 간의 평균값이 이용되고 있다. 캐나다 LFS에서는 6개월의 표본순환 방법을 이용하고 있기 때문에 3개월의 결과를 평균하게되면 결과적으로 1/3표본 수를 늘리는 효과를 갖게 된다. 만약 조사시점 간에 표본들이 정확히 일치한다면 추정의 효율 측면에서 이러한 이득은 기대할 수 없다.

연속조사에서 추정량의 정확도를 향상시키기 위해 서로 다른 시점에서 조사된 몇 개의 조사 자료를 풀링(pooling)하는 경우를 흔히 볼 수 있다. 특히 시점이 다른 몇 차례의 조사결과를 결합하거나 이들의 평균을 구하는 것이 보통의 방법인데, 이러한 방법은 소지역 통계 작성 시 해당 소지역에 배당된 표본 크기가 매우 적어서 추정의 정확도가 크게 떨어지는 경우에 유용하다. 그러나 다른 시점의 조사결과들을 결합하여 산출되는 추정값은 개념상의 문제점을 항상 내재하고 있다.

대영역 내에서 추정된 소지역 추정값들의 합계는 대영역의 추정값과 일치해야 하지만, 대개의 경우 소지역 추정값들의 합계는 대영역의 추정값과

일치하지는 않는다. 따라서 최종적으로 총계를 일치시키는 다음과 같은 보정이 이루어져야 한다. 소지역 i에 대한 총계 Y_i의 추정량을 $\hat{Y}_i$, 해당 소지역을 포함하는 대영역 a에 대한 $\hat{Y}_i$의 합계를 $\hat{Y}(a)$라 하자. 이때 소지역 추정값들은 다음 (3.63)식의 $\hat{Y}_i^{ADJ}$를 통해 보정된다.

$$\hat{Y}_i^{ADJ} = \frac{\hat{Y}_i}{\hat{Y}(a)} Y(a) , \qquad (3.63)$$

여기에서 $\hat{Y}(a) = \sum_{i \in a} \hat{Y}_i$ 이다.

(6) 노동력 추정값의 분산추정

(가) 배 경

캐나다 노동력 조사(LFS)는 캐나다 통계청에서 실시하는 가장 큰 규모의 월 단위 가구 조사로써 주로 전국 단위, 주 단위 및 주 내의 소지역 단위에 대한 다양한 노동력 특성에 대한 추정값들을 생산하고 있다. 캐나다의 LFS는 6개의 순환 패널을 갖는 층화 다단계 연동교체 표본설계를 따른다. 매번 인구 센서스 후 LFS는 표본설계에서 부분적인 보완이 이루어져 왔으며, 특히, 1981년에는 표본추출, 데이터 수집 및 추정 방법론 등에서 포괄적인 보완이 이루어졌다. 이 시기에 주 내의 소지역에 대한 추정치의 신뢰도를 높이기 위한 사후층화 비추정 절차가 새로이 마련되었다.

이 논문은 분산추정의 방법론에 대한 연구결과를 요약하였다. 과거 LFS의 분산추정은 Keyfitz 절차를 일반화한 Woodruff의 계산법이 이용되었다(Woodruff 1971). 이 방법은 Platek and Singh(1976)의 논문에서는 Keyfitz 방법으로 불리운다.

LFS에서는 다음 소개되는 세가지 유형의 지역들이 표본설계에 반영된다. 주요도시들로 구성되어 있는 SR지역(self-representing area), 소규모 도시들과 시골을 포함하는 NSR지역(non-self-representing area)과 군사지역 등과

같은 특수지역들이 이러한 유형의 지역들이다. NSR지역과 특수지역들에 대한 분산추정은 비추정방법에 Keyfitz 방법을 혼합하여 적용하였다. 이 단계 랜덤그룹 표본설계가 반영된 SR지역들에 대해서는 Rao, Hartley and Cochran(1962)과 Rao(1975)의 분산 추정량을 이용하였고, 이 방법을 Keyfitz 방법을 이용한 분산추정량과 비교하였다. 한편, 추정값들에 대한 분산 추정량들을 비 보정(ratio adjustment) 분산추정량들과 편향 및 안정성의 측면에서 비교하였다. 또한 반복 수의 증가에 따른 Keyfitz 분산추정량의 영향도 살펴보았다. 결론적으로 SR 지역에 대해서는 Keyfitz 방법이 훌륭한 대안으로 채택되었다.

(나) SR 설계에 대한 분산추정

① SR 설계(SR Design)

LFS 표본설계에서 SR 지역들은 이단계 랜덤그룹 설계방식을 취하며, 일단계 추출단위(PSU)들은 확률비례추출로 추출되며 이단계 추출단위들은 계통추출이 이루어진다. 하나의 층에 N개의 일차추출단위(PSU)가 있고, j번째 PSU에 대한 크기 측도를 x_j, $j=1,2,\cdots,N$, 거주단위의 수를 M_j, 층에서의 추출률을 $1/W$, 층으로부터 추출된 PSU의 수를 n이라 하자. N개의 PSU는 n개의 그룹으로 랜덤하게 분할되고 i번째 랜덤 그룹은 N_i개의 PSU들을 포함한다. 여기에서 $\sum_{i=1}^{n} N_i = N$ 이다. 우선 다음의 수식을 정의하자.

$$p_j = \frac{x_j}{\sum_{t=1}^{N} x_t} \quad , \quad j=1,2,\cdots,N_t \ ,$$

$$\delta_{ij} = \begin{cases} 1 \ , & \text{if the } j\text{th PSU is in the } i\text{th group} \\ 0 \ , & \text{otherwise} \end{cases}$$

이때 $\pi_j = \sum_{j=1}^{N} \delta_{ij} p_j$는 i번째 랜덤 그룹의 상대적인 크기를 나타낸다.

$a_{ij} = \delta_{ij} W p_j / \pi_{ij}$, $r_{ij} = a_{ij} - [a_{ij}]$이라 하고, $\{r_{ij}, j=1,2,\cdots,N\}$가 내림차순으

로 정렬되었다고 할 때, 계통추출의 추출간격 W_{ij}는 다음과 같이 정의할 수 있다.

$$W_{ij}=[a_{ij}]+1, \quad j=1,2,\cdots,R$$

$$=[a_{ij}], \quad j=R+1,\cdots,N,$$

여기에서 $R=\sum_{j=1}^{N} r_{ij}$, $\sum_{j=1}^{N} W_{ij}=W, \ i=1,2,\cdots,n.$

하나의 PSU는 n개의 랜덤 그룹 각각으로부터 추출률 W_{ij}에 비례하는 확률로 추출된다. i번째 랜덤 그룹으로부터 추출된 j번째 PSU는 $1/W_{ij}$의 비율로 부차추출된다. 이때 전체 추출률은 $1/W$ 이 된다. 각각의 랜덤 그룹은 1에서 6까지의 패널로 할당되고, 랜덤 그룹의 수 n은 보통 6의 배수이고 각각의 패널은 같은 수의 랜덤 그룹을 갖는다. 각 랜덤 그룹으로부터 하나의 PSU가 추출되므로 i번째 랜덤 그룹으로부터 추출된 PSU에서 부차추출률은 $1/W_i$이 된다. 랜덤 그룹 i에서 선택된 거주단위의 수는 m_i로 나타내기로 한다.

② 분산 추정량(Variance Estimator)

하나의 층에 대한 특성치 y의 총계에 관심이 있다고 하자. j번째 PSU에서 k번째 거주단위에 대한 y값을 y_{ik} ($k=1,2,\cdots,M_j$)라 할 때, 총계 $Y=\sum_{j=1}^{N}\sum_{k=1}^{M_i} y_{ik}$는 $\hat{Y}=W\sum_{i=1}^{n} y_i$ 로 추정될 수 있다. 여기에서 y_i는 i번째 그룹에서 선택된 PSU로부터 추출된 m_i개의 거주단위에 대한 y값들의 합을 나타낸다. $\hat{Y}$의 분산추정량은 다음과 같은 방법으로 추정될 수 있다.

Keyfitz의 분산 추정량(1957)

과거 표본설계에서 이용하였던 총계 추정량에 대한 분산 추정공식은 다음 (3.64)식과 같다.

$$\hat{V}_1(\bar{Y}) = W^2(\sum_o y_i - \sum_e y_i)^2 \ , \qquad (3.64)$$

여기에서 $\sum_o$는 홀수 패널들에 대한 합, $\sum_e$는 짝수 패널들에 대한 합을 나타낸다. 위의 (3.64)식을 일반화시킨 일반화 Keyfitz 분산 추정공식은 다음 (3.65)식과 같이 주어진다.

$$\hat{V}_2(\bar{Y}) = W^2 \frac{n}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \ , \qquad (3.65)$$

여기에서 $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ 이며, $\hat{V}_2(\bar{Y})$가 효율성이나 안정성 측면에서 $\hat{V}_1(\bar{Y})$보다 선호될 수 있다.

Rao, Hartley and Cochran의 분산 추정량(1962)

Rao, Hartley and Cochran의 분산 추정공식은 i번째 그룹으로부터 추출된 m_i개의 거주단위의 수가 고정되어있고, 거주단위들은 단순임의추출이 가정된 상태에서 유도된다. 분산 추정공식은 다음과 같이 주어진다.

$$\hat{V}_3(\bar{Y}) = A\sum_1^n \pi_i \left(\frac{M_i}{m_i} \frac{y_i}{p_i} - \bar{Y} \right)^2 + \sum_i^n \frac{\pi_i}{p_i} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i} \right) s_i^2 \ , \qquad (3.66)$$

여기에서 $A = \dfrac{\sum_1^n N_i^2 - N}{N^2 - \sum_1^n N_i^2}$, $s_i^2 = \dfrac{1}{m_i - 1} \sum_{k=1}^{m_i} (y_{ik} - \bar{y}_i)^2$. M_i는 i번째 그룹에서 선택된 PSU에 속해있는 거주단위들의 수이고 M_i개의 거주단위들 중 m_i개의 거주단위들이 계통추출로 추출되나 분산추정값은 단순임의추출 하에서 계산된다. i번째 그룹에서 선택된 PSU로부터 추출된 k번째 거주단위에 대한 y값이 y_{ik}이며 $\bar{y}_i$는 $\bar{y}_i = y_i / m_i$로 주어진다.

$\pi_i / p_i = W / W_i$ 이고, $M_i / m_i = W_i$ 이므로 (3.65)식의 분산 추정공식은 다음 (3.67)식과 같이 주어질 수 있다.

$$\hat{V}_3(\hat{Y}) = A\sum_1^n \pi_i \left(W\frac{y_i}{\pi_i} - \hat{Y} \right)^2 + W\sum_1^n \left(1 - \frac{m_i}{M_i} \right) M_i s_i^2 \qquad (3.67)$$

Rao의 분산 추정량(1975)

Rao의 분산 추정공식에서는 m_i 개의 거주단위들이 단순임의추출로 추출되나, 표본크기 m_i 를 확률변수로 취급하여 분산 추정공식을 유도하였다. Rao의 분산 추정공식은 다음 (3.68)식과 같이 주어진다.

$$\hat{V}_4(\hat{Y}) = A\sum_1^n \pi_i \left(W\frac{y_i}{\pi_i} - \hat{Y} \right)^2 + \sum_1^n \left\{ \frac{\pi_i^2}{p_i^2} - A\left(\frac{\pi_i}{p_i^2} - \frac{\pi_i^2}{p_i^2} \right) \right\} \frac{M_i^2 s_i^2}{m_i} - \sum_1^n \frac{\pi_i}{p_i} M_i s_i^2 \quad (3.68)$$

$$= \hat{V}_3(\hat{Y}) + W^2 \sum_1^n m_i s_i^2 \left\{ \left(1 - \frac{W_i}{W} \right) - A\left(\frac{1}{\pi_i} - 1 \right) \right\}$$

Rao의 분산 추정공식은 이 단계 표본추출에서 확률표본크기가 가정되기 때문에 음의 값이 나올 가능성이 있음에 주의해야한다.

③ Monte Carlo Study

네 가지의 분산 추정량들의 편향과 상대적인 안정성을 검토하기 위한 몬테카를로 연구가 수행되었다. 이용된 자료는 1981년 센서스 자료 중 조사지역 내의 약 20% 의 계통추출 표본이며 Halifax의 CMA(Census Metropolitan Area) 지역으로부터 19개의 층에 대해 검토가 이루어졌다. 추출률 $1/W$ 은 0.04 로 주어진다. 19 개 층에 대한 PSU의 수, 추출된 PSU의 수, 거주단위들의 수 및 기대 표본크기가 <표 3.18>에 주어졌다.

<표 3.18> 몬테카를로 연구를 위해 이용된 층

층	거주단위 수	PSU 수	추출된 PSU 수	기대 표본크기
1	737	49	6	29.5
2	490	33	4	19.6
3	745	45	6	29.8
4	720	34	6	28.8
5	621	37	6	24.8
6	630	38	6	25.2
7	503	31	4	20.1
8	340	23	4	13.6
9	472	33	4	18.9
10	468	33	4	18.7
11	367	28	4	14.7
12	390	23	4	15.6
13	626	36	6	25.0
14	650	39	6	26.0
15	350	22	4	14.0
16	736	46	6	29.4
17	573	35	6	22.9
18	773	48	6	30.9
19	866	64	8	34.6
합계	11,056	697	100	442.3

몬테카를로 기법을 이용하여 각 층에서 1,000개의 표본이 독립적으로 생성되었다. t번째 몬테카를로 표본생성으로부터 층 h에 대한 총계 Y_h의 추정값을 $\hat{Y}_{ht}$ ($h=1,2,\cdots,19$, $t=1,2,\cdots,1000$), $\hat{V}_{jht}$ ($j=1,2,3,4$)를 $\hat{Y}_{ht}$의 네 개의 분산추정량이라 하고 다음을 정의하였다.

$$Y=\sum_{h=1}^{19} Y_h \ ,$$

$$\hat{Y}_t=\sum_{h=1}^{19} \hat{Y}_{ht}, \quad t=1,2,\cdots,1000,$$

$$\hat{V}_{jt}=\sum_{h=1}^{19} \hat{V}_{jht}, \quad j=1,2,3,4 \ ,$$

여기에서 $\hat{Y}_t$는 t번째 몬테카를로 표본생성에서 얻어지는 총계 Y의 추정

값, $\hat{V}_{jt}$ ($j=1,2,3,4$)는 분산 추정값을 나타낸다.

몬테카를로 기대값과 분산을 각각 E^*와 V^*로 표기할 때 T개의 몬테카를로 표본생성에 대한 기대값과 분산은 각각 다음과 같이 주어진다.

$$E^*(\hat{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \hat{\theta}_t ,$$

$$V^*(\hat{\theta}) = \frac{1}{T}\sum_{t=1}^{T} [\ \hat{\theta}_t - E^*(\hat{\theta})]^2 .$$

위의 정의를 이용하여 $\hat{V}$의 몬테카를로 분산 $V^*(\hat{V})$과 분산추정량 $\hat{V}_j$의 몬테카를로 기대값 $E^*(\ \hat{V}_j)$와 몬테카를로 분산 $V^*(\ \hat{V}_j)$를 얻을 수 있다.

분산추정량 $\hat{V}_j$의 편향과 백분위 편향은 각각 다음과 같이 정의될 수 있다.

$$B_j = E^*(\ \hat{V}_j) - V^*(\hat{V}) ,$$

$$PB_j = 100\frac{B_j}{V^*(\hat{V})} ,\ j=1,2,3,4 .$$

이때 $\hat{V}_j$의 평균제곱오차 MSE는 다음과 같이 주어진다.

$$MSE_j = V^*(\ \hat{V}_j) + B_j^2 ,\ j=1,2,3,4 .$$

Keyfitz 분산추정량 $\hat{V}_1$에 대한 $\hat{V}_j$의 상대효율은 다음과 같이 정의할 수 있다.

$$\text{Rel. Eff}(\ \hat{V}_j \text{ vs. } \hat{V}_1) = (MSE_1 / MSE_j)^{1/2} ,\ j=2,3,4 .$$

분산추정량들에 대한 상대편향과 효율에 대한 결과는 <표3.19>과 <표3.20>에 주어졌다.

<표 3.19> 총계 추정값에 대한 분산추정량들의 백분위 편향

구 분	백분위 편향(PB_j)			
	$\hat{V}_1$	$\hat{V}_2$	$\hat{V}_3$	$\hat{V}_4$
취업인구	23.4	24.5	-4.7	-6.3
실업인구	6.3	6.6	3.7	1.2
노동력인구	24.2	25.2	-5.1	-6.7

<표 3.20> $\hat{V}_1$ 에 대한 $\hat{V}_2$, $\hat{V}_3$, $\hat{V}_4$ 의 상대효율

구 분	백분위 편향(PB_j)		
	$\hat{V}_2$	$\hat{V}_3$	$\hat{V}_4$
취업인구	1.51	3.22	3.11
실업인구	1.52	1.71	1.76
노동력인구	1.49	3.24	3.12

편향에 대해서는 분산추정량 $\hat{V}_1$ 과 $\hat{V}_2$ 가 유사하며, $\hat{V}_3$ 와 $\hat{V}_4$ 도 비슷한 결과를 보인다. 또한 $\hat{V}_1$ 과 $\hat{V}_2$ 의 편향은 취업인구와 노동력인구에서 비교적 큰 양의 편향값을 보이는 반면 $\hat{V}_3$ 와 $\hat{V}_4$ 의 편향은 상대적으로 작은 값을 나타낸다. 효율성 측면에서 $\hat{V}_3$ 와 $\hat{V}_4$ 가 서로 유사하며 $\hat{V}_1$ 과 $\hat{V}_2$ 에 비해서는 월등한 효율을 보인다. $\hat{V}_1$ 과 $\hat{V}_2$ 중에서는 $\hat{V}_2$ 의 효율이 더 좋게 나타난다.

전체인구에 대한 비 추정값들의 분산추정량에 대한 결과가 다음 <표 3.21>과 <표3.22>에 주어졌다. 비 추정값들에 대한 분산추정량은 $\hat{V}_j^{(R)}$ ($j=1,2,3,4$)로 표기하였다.

<표 3.21> 총계 추정값에 대한 분산추정량들의 백분위 편향

구 분	백분위 편향(PB_j)			
	$\hat{V}_1^{(R)}$	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
취업인구	3.7	4.3	-1.1	-3.1
실업인구	5.3	5.5	4.0	1.4
노동력인구	4.5	5.0	-0.5	-2.5

<표 3.22> $\hat{V}_1^{(R)}$ 에 대한 $\hat{V}_2^{(R)}$, $\hat{V}_3^{(R)}$, $\hat{V}_4^{(R)}$ 의 상대효율

구 분	백분위 편향(PB_j)		
	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
취업인구	2.13	2.59	2.52
실업인구	1.57	1.71	1.76
노동력인구	2.08	2.56	2.51

$\hat{V}_1^{(R)}$ 과 $\hat{V}_2^{(R)}$ 의 편향이 취업인구와 노동력인구에서 $\hat{V}_1$ 과 $\hat{V}_2$ 에 비해 훨씬 작아진 사실을 확인할 수 있다. $\hat{V}_3^{(R)}$ 과 $\hat{V}_4^{(R)}$ 의 편향도 취업인구와 노동력인구에서 $\hat{V}_3$ 와 $\hat{V}_4$ 보다 작은 값을 가지며 실업인구에서는 거의 변화가 발생하지 않았다.

몬테카를로 표본을 이용하여 네 가지 분산추정량들의 비 보정(ratio-adjustment) 추정값들에 대한 95% 신뢰구간을 살펴보았다. 추정값들의 95% 신뢰구간에 대한 포함비율이 다음 <표 3.23>에 주어졌다.

<표 3.23> 비보정을 갖는 총계 추정값들의 95% 신뢰구간 포함비율

구 분	포함 비율			
	$\hat{V}_1^{(R)}$	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
취업인구	93.6	95.4	94.6	94.2
실업인구	94.3	95.1	95.3	95.0
노동력인구	93.2	95.3	94.6	94.2

취업인구, 실업인구 및 노동력인구의 모든 부분에서 네 가지 분산추정량들의 수행결과가 모두 적합한 것으로 나타났다.

편향의 관점에서는 비 보정 추정값들의 분산추정량들이 서로 상이한 결과를 나타내지는 않는다. 상대효율에서는 $\hat{V}_3^{(R)}$와 $\hat{V}_4^{(R)}$가 $\hat{V}_2^{(R)}$에 비해 효율이 높은 것으로 나타났다. 한편, $\hat{V}_1^{(R)}$의 자유도는 19이고(각 층별로 1개의 자유도를 가짐) $\hat{V}_3^{(R)}$는 각 PSU가 하나의 반복으로 처리되어 81개의 자유도를 갖게 된다. 따라서 비 보정 추정값들에 대한 Keyfitz 분산 추정량은 반복 수를 증가시키면 추정량의 안정성을 확보할 수 있다.

④ 반복 수를 갖는 Keyfitz 분산 추정량

Keyfitz 방법의 효율성을 높이기 위해 6개의 순환 패널들이 반복표본들로 채택되었다. 6개의 순환 패널을 반복표본으로 이용한 분산 추정값들과 과거 표본설계를 이용한 2개의 반복표본으로부터 계산된 분산 추정값들이 비교되었다. 순환 패널이 반복표본으로 처리됨에 따라 기인된 중요한 관심사항은 패널 편향으로부터 발생할 수 있는 분산 추정값들의 증가부분이다. 이러한 부분을 살펴보기 위해 '85년 3월부터 '87년 2월까지의 24개월의 LFS 자료가 이용되었다. 취업인구, 실업인구, 노동력인구의 24개월에 대한 분산 추정값들에 대한 평균과 표준편차를 계산하였다. 2개의 반복표본과 6개의 반복표본 하에서 얻어진 분산들에 대한 평균과 표준편차의 비는 24개의

CMA(Census Metropolitan Area) 지역들의 평균으로 산출하였고 다음 <표 3.24>에 주어졌다.

<표 3.24> 단순임의 분산추정값의 비교(LFS의 CMA지역 자료)

구 분	분산의평균에 대한 비 평균(average ratio) : 2 vs. 6 반복	분산의 표준편차에 대한 비 평균(average ratio) : 2 vs. 6 반복
취업인구	0.997	1.813
실업인구	0.995	1.515
노동력인구	1.003	1.833

6개의 반복표본을 이용한 분산이 2개의 반복표본의 분산보다 작은 값을 나타낸다. 순환 패널을 반복으로 채택할 경우 분산 추정값들의 편향에 거의 영향을 미치지 않으며, 6개의 반복표본을 이용한 분산이 2개의 반복표본보다는 훨씬 안정적임을 확인할 수 있다. 즉 Keyfitz 방법에서 6개의 순환 패널을 반복으로 사용할 경우 심각한 편향은 발생하지 않으며 2개의 반복표본을 이용했을 경우보다는 효율이 증가됨을 확인할 수 있다.

(다) 비 추정값 탐색을 위한 분산추정방법

① LFS에서 비 추정방법

과거 LFS에서는 사후층화 비 추정방법이 이용되었다. 무응답을 보정하기 위한 일종의 설계 가중치인 부차 가중치가 LFS 목표 모집단의 추정치들에 대해 비 보정되었다. 이러한 비 추정방법은 주 지역의 특성치에 대한 추정의 신뢰도를 높이는 결과를 보였으나 주 내의 소지역들에 대해서는 문제점을 안고 있었다. 주 내의 ER(Economical Region) 지역과 CMA(Census Metropolitan Area) 지역들에 대한 추정의 정확도를 높이기 위해 탐색적인 비 추정 절차가 채택되었다.

탐색적 추정절차는 보정값들의 수열을 통해 수행된다. 먼저 부차가중치가

주 내의 소지역의 인구를 참조하여 보정되며, 이 후 성별-연령대별 범주를 반영한 주 수준의 보정값이 최종 가중치에 적용된다. 이러한 절차는 한번 더 수행되어 한쌍의 가중치가 추가적으로 생성된다. W_0를 부차 가중치라 하고 (W_1, W_2)와 (W_3, W_4)를 2회 반복으로부터 생성된 가중치들의 쌍이라 하자. 노동력 특성값들은 W_4를 이용하여 추정된다. 주 지역의 성별-연령대별 그룹들에서 주변 총계 W_4는 상응하는 그룹들의 외부 인구 추정치와 정확히 일치하나 주 내의 소지역에 대해서는 반드시 그렇지는 않다. 그러나 그 차이는 매우 작게 나타난다.

② 1회 반복 비 추정값에 대한 분산공식

1회 반복 비 추정값들에 대한 분산공식을 유도하면 다음과 같다. 여기에서 적용되는 기본적인 방법론은 선형적인 형태의 부차 가중치를 얻을 때까지 테일러 전개 근사식을 연속적으로 적용하는 방법이다. 세부적인 유도과정을 소개하면 다음과 같다.

$Y^{(0)}, Y^{(1)}, Y^{(2)}$를 주 지역에서 W_0, W_1, W_2에 근거하여 추정된 노동력 특성값 y의 추정값들이라 하자. 이때 $Y^{(2)}$는 다음 (3.69)식과 같이 주어질 수 있다.

$$Y^{(2)} = \sum_a \frac{Y_a^{(1)}}{P_a^{(1)}} P_a \ , \qquad (3.69)$$

여기에서 $Y_a^{(1)}$은 주 지역에서 성별-연령대별 그룹 a에 대한 특성치 y의 W_1가중추정값, $P_a^{(1)}$은 주 지역에서 성별-연령대별 그룹 a에 대한 인구의 W_1가중 추정값, P_a는 주 지역에서 성별-연령대별 그룹 a에 대한 인구의 외부 추정치를 나타낸다.

F_a를 $F_a = Y_a^{(1)} / P_a^{(1)}$이라 할 때, $(E(Y_a^{(1)}), E(P_a^{(1)}))$에서 F_a에 대한 일계 테일러 근사식을 구하면 다음과 같이 주어진다.

$$F_a \doteq \frac{E(Y_a^{(1)})}{E(P_a^{(1)})} + \frac{1}{E(P_a^{(1)})}\{Y_a^{(1)} - E(Y_a^{(1)})\} - \frac{E(Y_a^{(1)})}{\{E(P_a^{(1)})\}^2}\{P_a^{(1)} - E(P_a^{(1)})\}.$$

이때 $Y^{(2)}$의 분산에 대한 테일러 근사식은 다음 (3.70)식과 같이 주어질 수 있다.

$$V(Y^{(2)}) = V(\sum_a F_a P_a)$$

$$\doteq V\left\{\sum_a \frac{P_a}{E(P_a^{(1)})}(Y_a^{(1)} - R_{Y_a}^{(1)} P_a^{(1)})\right\}, \qquad (3.70)$$

여기에서 $R_{Y_a}^{(1)} = E(Y_a^{(1)})/E(P_a^{(1)})$을 나타낸다.

다음으로 W_1 가중 추정값 $Y_a^{(1)}$과 $P_a^{(1)}$은 W_0 가중 추정값의 항으로 다음 (3.71)식과 같이 나타낼 수 있다.

$$Y_a^{(1)} = \sum_s \frac{Y_{sa}^{(0)}}{P_s^{(0)}} P_s, \qquad (3.71)$$

$$P_a^{(1)} = \sum_s \frac{P_{sa}^{(0)}}{P_s^{(0)}} P_s,$$

여기에서 s는 CMA 또는 ER 지역을 나타내며, P_s는 지역 s의 인구를 나타낸다. (3.71)식의 $Y_a^{(1)}$과 $P_a^{(1)}$을 (3.70)식에 대입하여 W_0 가중 추정값의 비에 대한 일계 테일러 근사식을 구하면 다음 (3.72)식과 같이 주어질 수 있다.

$$V(Y^{(2)}) \doteq V\left[\sum_a \frac{P_a}{E(P_a^{(1)})} \sum_s \frac{P_s}{E(P_s^{(0)})}\{(Y_{sa}^{(0)} - R_{Y_{sa}}^{(0)} P_s^{(0)}) - R_{Y_a}^{(1)}(P_{sa}^{(0)} - R_{P_{sa}}^{(0)} P_s^{(0)})\}\right] \qquad (3.72)$$

여기에서 $R_{Y_{sa}}^{(0)} = E(Y_{sa}^{(0)})/E(P_s^{(0)})$, $R_{P_{sa}}^{(0)} = E(P_{sa}^{(0)})/E(P_s^{(0)})$이다.

위의 (3.72)식은 다음 (3.73)식과 같이 축약된 형태로 다시 표현할 수 있다.

$$V(Y^{(2)}) \doteq V\left\{\sum_s \sum_{h \in s} \sum_{i=1}^{n_h} \sum_a (Z_{Y_{hia}}^{(0)} - R_{Y_a}^{(1)} Z_{P_{hia}}^{(0)})\right\}$$

$$= V\left(\sum_{s} \sum_{h \in s} \sum_{i=1}^{n_h} D_{shi}^{(0)} \right) , \tag{3.73}$$

여기에서 $D_{shi}^{(0)} = \sum_{a} \left(Z_{Y_{shia}}^{(0)} - R_{Y_a}^{(1)} Z_{P_{shia}}^{(0)} \right)$,

$$Z_{Y_{shia}}^{(0)} = \frac{P_a}{E(P_a^{(1)})} \frac{P_s}{E(P_s^{(0)})} \left(Y_{shia}^{(0)} - R_{Y_{sa}}^{(0)} P_{shi}^{(0)} \right) ,$$

$$Z_{P_{shia}}^{(0)} = \frac{P_a}{E(P_a^{(1)})} \frac{P_s}{E(P_s^{(0)})} \left(P_{shia}^{(0)} - R_{P_{sa}}^{(0)} P_{shi}^{(0)} \right)$$

이고, h는 s에 속하는 층을 나타내며, i는 층 h에서 반복을 나타낸다.

식 (3.73)에서 $\left\{ \sum_{i=1}^{n_h} D_{shi}^{(0)} \right\}$는 부차가중치들에 의해 결정되므로 독립성을 가정할 수 있다. 따라서 식 (3.73)는 다음 (3.74)식과 같이 표현될 수 있다.

$$V(Y^{(2)}) \doteq V\left(\sum_{h \in s} \sum_{h} \sum_{i=1}^{n_h} D_{shi}^{(0)} \right)$$

$$= V\left(\sum_{h} \sum_{i=1}^{n_h} \sum_{s \ni h} D_{shi}^{(0)} \right) \tag{3.74}$$

여기에서 $\sum_{s \ni h}$는 층 h를 포함하고 있는 주 내의 모든 소지역들에 대한 합을 나타낸다. $D_{hi}^{(0)}$를 $D_{hi}^{(0)} = \sum_{s \ni h} D_{shi}^{(0)}$와 같이 정의하면, 위의 (3.74)식은 다음과 같이 주어진다.

$$V(Y^{(2)}) \doteq V\left(\sum_{h} \sum_{i=1}^{n_h} D_{hi}^{(0)} \right) \tag{3.75}$$

여기에서 $\left\{ \sum_{i} D_{hi}^{(0)} \right\}$는 부차가중치에 의해 결정되므로 이 변수들은 독립성을 가정할 수 있으며, 분산은 다음 식으로부터 추정될 수 있다.

$$\hat{V}(Y^{(2)}) \doteq \sum_{h} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (D_{hi}^{(0)} - \overline{D}_h^{(0)})^2 \tag{3.76}$$

여기에서 $\overline{D}_h^{(0)} = \frac{1}{n_h} \sum_{i=1}^{n_h} D_{hi}^{(0)}$이다. 그러나 이 표현식에서는 기대값이 포함되어있고 이러한 값들은 미지의 값이므로 추정값으로 대체하여 분산의 추정값

을 근사적으로 계산할 수 있으며 이를 이용한 분산 추정값은 최종적으로 다음 (3.77)식과 같이 주어진다.

$$\hat{V} \doteq \sum_h \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (D_{hi}^{(2)} - \overline{D}_h^{(2)})^2 \qquad (3.77)$$

여기에서 $D_{hi}^{(2)} = \sum_{s \in h} D_{shi}^{(2)}$,

$$\overline{D}_h^{(2)} = \frac{1}{n_h} \sum_{i=1}^{n_h} D_{hi}^{(2)},$$

$$D_{shi}^{(2)} = \sum_a (Z_{Y_{shia}}^{(2)} - R_{Y_s}^{(2)} Z_{P_{shia}}^{(2)}),$$

$$Z_{Y_{shia}}^{(2)} = \frac{P_a}{P_a^{(1)}} \frac{P_s}{P_s^{(0)}} \left(Y_{shia}^{(0)} - \frac{Y_{sa}^{(0)}}{P_s^{(0)}} P_{shi}^{(0)} \right) = Y_{shia}^{(2)} - \frac{P_{shi}^{(0)}}{P_s^{(0)}} Y_{sa}^{(2)},$$

$$Z_{P_{shia}}^{(2)} = \frac{P_a}{P_a^{(1)}} \frac{P_s}{P_s^{(0)}} \left(P_{shia}^{(0)} - \frac{P_{sa}^{(0)}}{P_s^{(0)}} P_{shi}^{(0)} \right) = P_{shia}^{(2)} - \frac{P_{shi}^{(0)}}{P_s^{(0)}} P_{sa}^{(2)},$$

$$R_{Y_s}^{(2)} = \frac{Y_a^{(1)}}{P_a^{(1)}} = \frac{P_a}{P_a^{(1)}} \frac{Y_a^{(1)}}{P_a} = \frac{Y_a^{(2)}}{P_a}.$$

위의 분산 추정공식 (3.77)식은 노동력 특성값들의 W_2 가중 추정값들에 대한 추정공식이며 W_0와 W_2의 두 가중치의 값을 요구한다.

③ 2회 반복 비 추정값의 분산 추정

2회 반복 비 추정값에 대한 분산공식은 3.2절을 응용하여 테일러 급수전개의 연속적인 적용으로 얻을 수 있다. 그러나 2회 반복에 기인한 분산 추정공식은 매우 복잡한 형태를 취하기 때문에 1회 반복에 기인한 분산 추정공식이 오히려 합리적일 수 있다. 1회 반복 분산 공식은 한 쌍의 가중치 (W_0, W_2)를 이용한다. 여기에서는 (W_0, W_2) 대신에 (W_0, W_4)를 이용하였다. W_2보다는 W_4가 노동력 추정값들의 CV값들에 강한 영향력을 주지 않기 때문이다. 주 지역인 Nova Scotia 지역의 1981년 센서스 자료를 이용하여 몬테카를로 시뮬레이션이 수행되었다. 각각의 몬테카를로 표본에서 LFS 표본설계가 매 단계의 표본 추출을 통해 검증되었다. 1,000개의 몬테카를로

표본들이 독립적으로 추출되었다. 각각의 몬테카를로 표본들에 대해 2회 반복표본 비 추정값($Y^{(4)}$), 1회 반복 분산 추정량과 이의 CV 추정값을 이용한 분산 추정값($\hat{V}(Y^{(4)})$)과 95% 신뢰구간($Y^{(4)} \pm 1.96\sqrt{\hat{V}(Y^{(4)})}$)을 주 지역과 주 지역내의 소지역들에 대해 계산하였다. 또한 1,000개의 CV 값들의 평균을 계산하여 실제값과 매우 유사한 몬테카를로 CV 값들과 비교하였다. 결과는 <표 3.25>에 주어졌다. <표 3.25>의 모든 셀에 대해서 CV 값의 차이는 8% 미만으로 나타났고, 21개의 셀 중 13개의 셀에서는 4%미만의 값을 갖는다. 한편 신뢰구간의 포함범위는 <표 3.26>에 주어졌다. 취업인구와 노동력인구에 대한 95% 신뢰구간의 포함범위는 만족한 값을 보이나 실업인구에 대해서는 다소 낮은 포함범위를 나타내나 여전히 받아들일만한 결과를 보인다.

<표 3.25> 1회 반복 분산추정량의 평균 CV값과 몬테카를로 CV값

구 분	ER 210	ER 220	ER 230	ER 240	ER 250	CMA Halifax	Province (Nova Scotia)
				Average CV's			
취업인구	3.52	3.46	3.14	3.05	1.96	2.01	1.08
실업인구	10.36	12.28	13.13	13.43	10.35	10.55	5.27
노동력인구	2.98	3.17	2.85	2.73	1.77	1.83	0.91
				Monte Carlo CV's			
취업인구	3.48	3.35	2.95	2.86	1.97	1.99	1.11
실업인구	10.90	12.71	13.28	13.37	11.12	11.31	5.59
노동력인구	2.76	3.08	2.76	2.53	1.72	1.74	0.92

<표 3.26> 1회 반복 분산추정량에 의한 95% 신뢰구간의 포함범위

구 분	ER 210	ER 220	ER 230	ER 240	ER 250	CMA Halifax	Province (Nova Scotia)
취업인구	94.5	92.8	94.0	94.7	94.7	94.9	92.5
실업인구	92.1	90.7	91.4	91.8	92.7	92.7	93.1
노동력인구	96.2	93.0	93.6	95.2	95.2	96.0	94.0

(라) 결 론

비 보정을 하지 않은 추정값들의 Keyfitz 분산 추정법은 매우 큰 양의 편향을 가지며 효율성도 크게 떨어진다. 반면에 탐색적 비 보정 추정방법은 상대적으로 작은 편향을 가지며 효율성도 크게 향상되는 것으로 확인되었다. 이 논문에서 소개된 비 보정 추정값들에 대한 분산 추정방법들은 무시할 수 있을 정도의 작은 편향을 갖는다. 한편 Keyfitz 방법은 반복 수를 증가시킬 경우 다른 분산 추정방법에 비해 효율을 크게 향상시킬 수 있었다. LFS 자료에서 6개의 순환 패널을 반복으로 취급하여 Keyfitz 방법을 적용시켜본 결과 순환 패널 편향에 기인한 Keyfitz 추정분산의 편향은 발생하지 않았다. Keyfitz 방법에 의해 유도된 1회 반복 분산공식은 2회 반복의 탐색적 비 추정값들에 대해서 매우 합리적인 분산 추정값들을 제공하며 신뢰구간에 대한 포함범위도 좋은 특성을 나타냈다.

3.4 프랑스

노동력 조사로부터 국가 수준의 노동 통계는 일년에 한 번씩 발표하여 이용자들에게 불편을 주었으나, 지역에서 실업 통계를 생산할 때 일정한 오차 범위 내로 기준을 충족하도록 하는 소지역 추정법에 관한 연구가 진행되고 있다.

표본 설계는 지역에 의해서 층화하였으나 실업자 수 또는 실업률 등의 집

적된 자료의 이용시에는 무응답을 보정하여 국가 수준으로 통계를 작성한 후에 성별-연령으로 분할했으므로 지역 통계의 신뢰성에 대해서 유의해야 할 것이다.

프랑스 통계청에서 취업과 실업에 관한 지역 통계의 생산을 위해서 노동력 조사와 센서스 뿐 만 아니라 실업 보험 신청자료 등 행정 업무 자료를 이용하는 체계를 발전시켜왔으나 소지역이나 세분화된 범주의 통계 작성은 미흡하다. 특히 취업 통계는 연말에 각 회사로부터 취업자 수에 대한 자료를 수집하고 있으나 작년말 기준으로 변화율에 대한 자료로 활용하여 국가 수준에서 비율에 대한 통계를 작성하고 있으며 실업자의 수에 대한 직접 통계를 작성하지는 않는다. 그래서 취업자 통계는 실업 보험과 센서스 자료 등이 핵심적인 역할을 하고 있다. 취업 통계는 해당 익년에 이용가능하다.
실업 통계는 분기별로 지역 통계와 국가 통계를 동시에 발표하며, 주로 이용되는 자료는 노동력 조사와 구직 등록자의 수이다. ILO 기준의 실업자와 실제 구직 등록자 및 노동력 조사에서 실업자와 차이를 반영하기 위해서 노동력 조사에서 추정된 ILO 실업자와 구직 등록자 수의 국가적 비율을 추정하여 활용하고 있다.

경제활동인구는 실업자와 취업자를 합해서 추정하지만 취업자는 직장을 갖고 있는 사람과 가사를 돌보는 사람을 합산해야 하므로 매 해마다 연말에 외삽법으로 조정하여 분기별 통계를 수정하여 시계열 통계로 관리한다.
취업자 통계에서 문제점은 연말에 센서스 자료를 이용하고 있으나 센서스 간격이 너무 길어서 인구 변동과 상황 변화를 제대로 반영할 수 없다는 점이다 (센서스:1968, 1975, 1982, 1990, 1999). 특히, 지역 취업 통계 작성 시에는 더 심각해질 수 있다.

실업 통계 작성에서 문제점은 노동력 조사에서 국가 단위의 ILO 실업 통계는 작성할 수 있으나 지역 단위 또는 소지역 단위의 실업 통계는 생산할

수 없으며, 구조적인 특성(인구 사회학적 구성)이 각 지역이 국가와 동일하다는 전제에서 계산되며 표준 오차를 계산하지 못하고 있다.

취업 통계 작성의 취약점을 보완하기 위해서 "ESTEL"이라는 프로젝트가 수년간 추진 중에 있으나 아직 1996년 통계를 시험 중에 있다. 특징은 국가와 지역의 취업 통계의 질과 신뢰도를 제고하는데 있으며 매년 말을 기준으로 하여 조사 자료와 행정 자료를 반영하고 있다. 센서스 자료를 기준으로 반영하는 것은 센서스 해에만 적용하고 다른 해에는 연말의 행정 업무와 조사 자료를 기준으로 삼는 것이 큰 차이점이다.

핵심적인 과제는 경제 활동 관련 통계의 기준은 센서스 정보이므로 센서스 실시를 일정한 간격으로 실시하는 계획일 것이다. 매 5년마다 센서스에 준하는 대규모 통계 조사 계획을 추진하고 있다.

제 4장 광주광역시와 충청북도의 실증조사

4.1 조사 목적

2000년 11월 1일 기준으로 실시한 인구주택 총 조사의 자료를 이용하여 2002년 경제활동인구조사의 표본설계의 개편안을 연구할때 시군구 단위 소지역 실업통계를 생산할 수 있는 방안을 연구하는 기본방향을 제공하고자 한다.

시군구 단위의 실업통계 작성을 위해서 무한정으로 표본조사구를 늘릴 수 없으므로 현재의 표본가구 규모 수준에서 시군구 실업통계를 일정 수준의 정확도를 유지하면서 생산할 수 있는 소지역 추정법의 연구가 2000년 6월부터 시작되었다.

그 동안 학계와 통계청의 공동연구로 소지역 추정법의 기본이론 요약, 통계 선진국인 미국, 캐나다와 영국 등의 소지역 실업통계 작성에서 소지역 추정법의 활용 실태와 경제활동인구조사 자료를 이용한 충북 시군의 실업통계 작성 연구 등이 추진되었다. 이제까지 연구된 간접 추계법인 복합 추정법에 의해서 경제활동인구조사 자료와 상주인구 추계자료를 이용하여 시군구의 실업통계를 작성할 수 있다는 가능성을 검증하고 연구 결과의 적용상의 문제점은 없는가를 분석하기 위해서 광주광역시와 충청북도에서 현재의 경제활동인구조사를 위한 표본조사구 외에 추가로 88개 조사구를 조사하고자 한다.

실증조사의 주된 목적은 기존의 표본조사구들의 경제활동인구조사 자료만을 이용하여 복합 추정법으로 시군구의 실업자를 추계한 추정값의 신뢰성과 정확도를 평가하기 위해서 별도로 추가된 88개 조사구의 자료와 기존의

151개 조사구의 자료를 통합하여 시군구의 실업자를 기존 방법(현 통계청 추정법)으로 산정한 추정값을 비교 분석하여 복합 추정법의 타당성을 검증할 뿐만 아니라 좀 더 효율적인 시군구 실업통계 작성 모형을 개발하는데 있다.

4.2 조사 개요

(1) 조사 기간 및 규모

연구 기간을 고려하여 조사 기간을 2001년 5월과 6월로 하였으며 조사주기도 경제활동인구조사와 같이 매월 15일이 속한 주간을 조사대상 기간으로 하였다. 조사지역은 광주광역시와 충청북도로 하였으며 실증조사에 사용될 추가된 조사구 수의 결정은 광주광역시는 기존 조사구가 87개 있으므로 5개 조사구의 평균이 20개 조사구가 될 수 있도록 14개 조사구를 추가하여 101개 조사구로 하였으며, 충청북도는 11개 시군에 64개 기존의 표본조사구가 있어서 74개 조사구를 추가하여 시군별로 최소한 10개의 조사구를 조사하여 어느 정도의 신뢰성을 갖는 실업통계를 생산할 수 있게 하였다.

광주광역시와 충청북도의 각 시군구별 주민등록인구와 표본조사구의 배분현황은 다음 <표4.1>에 요약되었다.

<표 4.1> 시군구별 조사구 분포와 주민등록인구

시도	시군구	조사구 수			주민등록인구
		기존	추가	합계	
광주 광역시	동구	10	4	14	123,647
	서구	15	3	18	287,867
	남구	16	2	18	231,501
	북구	34	0	34	475,992
	광산구	12	5	17	252,902
	소계	87	14	101	1,371,909
충청 북도	청주시	22	0	22	580,861
	청원군	5	9	14	123,984
	진천군	1	9	10	60,121
	괴산군	5	6	11	76,666
	충주시	11	5	16	217,305
	제천시	4	10	14	147,950
	음성군	6	5	11	87,956
	단양군	2	8	10	40,000
	옥천군	3	7	10	60,798
	보은군	2	8	10	43,245
	영동군	3	7	10	58,627
	소계	64	74	138	1,497,513
합계		151	88	239	

(2) 표본조사구 추출 및 조사방법

추가된 조사구는 1997년 경제활동인구조사 표본설계 시에 표본 추출틀로 사용한 '95년 인구주택 총 조사의 10% 표본조사구들 중에서 각 시군구별로 현재 표본조사구로 사용 중인 것을 제외한 후에 각 조사구들의 가구수에 비례하는 확률계통추출법으로 선정하였다. 선정된 추가 표본조사구 중에서

적합한지 여부는 각 시군구별로 지방 통계사무소에서 확인 후에 추가된 조사구로 확정하였다.

조사방법은 기존의 표본조사구는 2001년 5월과 6월의 경제활동인구조사 결과의 자료를 활용하기 위해서 전산 입력된 자료를 사용하고 추가 조사구에 대해서는 2개 조사구 당 한명의 임시 조사원을 채용하여 실증조사 목적 및 의의, 면접 및 현장조사 기법, 조사구 요도 작성 및 가구명부 작성법, 조사표 작성법과 조사 사항 내검요령 등을 교육하고 숙지시킨 후에 2개월간 조사를 담당케 하였다.

4.3 본 조사 및 자료 입력

임시 조사원은 조사기법과 조사업무 수행절차에 대해서 기본교육을 받은 후에 조사구 확인, 보완 및 요도를 작성한 후에 준비조사, 본 조사와 조사표 내검 및 보완과정을 통해서 본 조사를 수행하였다.

(1) 기본 교육

통계청 조사관리과와 사회통계과 담당직원이 임시 조사원들에게 시범실습식 집체교육을 통해서 다음 내용을 숙지 시켰다.

- 실증 조사의 중요성과 목적
- 면접 및 현장조사 기법과 대응 요령
- 조사구 요도 작성 및 가구명부 작성방법
- 조사 지침서 및 조사표 기입방법
- 조사 내용 내검 및 자료 입력방법

(2) 준비 조사

본 조사를 실시하기 전에 응답률을 높이고 조사의 정확성을 높이기 위해서 현장 확인 및 조사 대상자의 접촉 가능성을 확인하는 과정으로 아래와 같은 내용의 업무를 수행한다.

- 조사구 현장 방문
- 조사구의 유고 및 조사 가능 상태 확인 및 대체
- 조사구 경계 확인 및 거처 변동사항 파악
- 조사구 요도 보완 및 가구명부 작성
- 조사구역 선정
- 표본가구 확정(조사구 당 24가구)
- 조사구 최종 확인 및 교체
- 조사구 요도 및 가구명부 작성, 조사구역 및 대상가구 확정

(3) 본 조사

임시 조사원이 조사대상 가구를 직접 방문하여 친숙도를 넓힌 후에 정확한 자료를 수집하는 과정으로 아래와 같은 사항의 업무를 수행한다.

- 조사대상 가구 방문, 조사 홍보용 전단 및 답례품 전달
- 보조 조사표 배포 및 회수
- 응답자 면접 및 조사표 작성
- 조사내용 검토
- 조사표 내검 및 전산 입력

(4) 자료 입력

임시로 채용한 내검 및 입력요원이 조사된 조사표의 내용을 내검한 후

에 해당 사무소(출장소)에서 컴퓨터 입력 후 입력된 내용의 오류를 확인하고 입력된 자료를 송부한다. 조사표를 전산입력 완료 후에 조사관리과로 송부하며 조사입력은 조사 실시 후 10일 이내에 완료한다.

자료입력 후 송부된 전산자료에서 기존 조사구의 조사결과는 전산화일에서 광주광역시와 충북의 자료만 추출하고 추가 조사구의 조사결과는 Batch 방식의 별도 전산화일을 생산하여 두 개 파일을 통합한 후 실증조사 분석을 위한 자료화일을 만든다. 만들어진 자료에 대한 간단한 내용은 다음 절에서 살펴보자.

4.4 조사자료 요약

광주광역시에서 기존 조사구 87개와 추가 조사구 14개를 합한 101개 조사구에 대해서 남·여별 경제활동인구, 취업자, 실업자와 비경제활동인구에 대한 원자료(Raw Data)는 <부록1>에 수록하여 앞으로 심층적인 연구에서 참고자료로 활용토록 하였으며, 충청북도의 경우에도 기존 조사구 64개와 추가 조사구 74개를 합한 138개 조사구별로 남·여별 경제활동인구, 취업자, 실업자와 비경제활동인구에 대한 원자료를 <부록2>에 수록하였다.

광주광역시의 구별 조사된 자료의 요약은 <표4.2>와 <표4.3>에 주어졌다. 5월 자료에서 동구의 실업률은 기존 조사구에서는 4.56%, 추가 조사구에서는 5.82%이고 서구의 실업률은 기존 조사구에서는 4.97%이고 추가 조사구에서는 6.02%로 나타난다. 전체 지역에 대해서는 기존 조사구에서의 실업률은 4.29%, 추가 조사구에서의 실업률은 5.75%로써 추가 조사구에서의 실업률이 높게 나타난다.

6월 자료에서 동구의 기존 조사구의 실업률은 3.85%, 추가 조사구의 실업률은 4.90%이고 서구의 기존 조사구의 실업률은 6.43%, 추가 조사구의 실업률은 5.39%로 나타나며, 전체 지역들에 대해서는 기존 조사구의 실업률은 4.33%, 추가 조사구의 실업률은 5.80%로써 추가 조사구에 대한 실업률이 높게 나타난다.

<표4.2> 광주광역시 5월 구별 자료요약

구분		경제활동인구			취업자			실업자			비경제활동인구		
		기존	추가	합계	기존	추가	합계	기존	추가	합계	기존	추가	합계
동구	남자	141	63	204	130	59	189	11	4	15	99	44	143
	여자	144	40	184	142	38	180	2	2	4	145	73	218
	계	285	103	388	272	97	369	13	6	19	244	117	361
서구	남자	262	45	307	251	41	292	11	4	15	158	26	184
	여자	200	38	238	188	37	225	12	1	13	252	50	302
	계	462	83	545	439	78	517	23	5	28	410	76	486
남구	남자	255	38	293	241	33	274	14	5	19	130	16	146
	여자	234	26	260	223	23	246	11	3	14	233	36	269
	계	489	64	553	464	56	520	25	8	33	363	52	415
북구	남자	563	-	-	543	-	-	20	-	-	281	-	-
	여자	475	-	-	454	-	-	21	-	-	559	-	-
	계	1038	-	-	997	-	-	41	-	-	840	-	-
광산구	남자	241	84	325	233	82	315	8	2	10	64	53	117
	여자	167	66	233	162	64	226	5	2	7	151	85	236
	계	408	150	558	395	146	541	2	4	17	215	138	353
합계	남자	1462	230	1692	1398	215	1613	64	15	79	732	139	871
	여자	1220	170	1390	1169	162	1331	51	8	59	1340	244	1584
	계	2682	400	3082	2567	377	2944	115	23	138	2072	383	2455

<표4.3> 광주광역시 6월 구별 자료요약

구분		경제활동인구			취업자			실업자			비경제활동인구		
		기존	추가	합계	기존	추가	합계	기존	추가	합계	기존	추가	합계
동구	남자	140	63	203	132	58	190	8	5	13	104	44	148
	여자	146	39	185	143	39	182	3	0	3	143	74	217
	계	286	102	388	275	97	372	11	5	16	247	118	365
서구	남자	60	250	310	60	232	292	0	18	18	47	142	189
	여자	80	158	238	71	154	225	9	4	13	68	230	298
	계	140	408	548	131	386	517	9	22	31	115	372	487
남구	남자	294	201	495	277	191	468	17	10	27	158	254	412
	여자	260	266	526	246	241	487	14	25	39	279	151	430
	계	554	467	1021	523	432	955	31	35	66	437	405	842
북구	남자	565	-	565	543	-	543	22	-	22	279	-	279
	여자	477	-	477	456	-	456	21	-	21	560	-	560
	계	1042	-	1042	999	-	999	43	-	43	839	-	839
광산구	남자	251	83	334	243	82	325	8	1	9	66	56	122
	여자	176	60	236	172	58	230	4	2	6	148	90	238
	계	427	143	570	415	140	555	12	3	15	214	146	360
합계	남자	1310	597	1907	1255	563	1818	55	34	89	654	496	1150
	여자	1139	523	1662	1088	492	1580	51	31	82	1198	545	1743
	계	2449	1120	3569	2343	1055	3398	106	65	171	1852	1041	2893

충청북도의 시군별 조사된 자료의 요약은 <표4.4>와 <표4.5>에 주어졌다. 5월 자료에서 제천시의 실업률은 기존 조사구에서는 1.83%, 추가 조사구에서는 2.49%이고 충주시의 실업률은 기존 조사구에서는 2.28%이고 추가 조사구에서는 1.25%로 나타난다.

전체 지역에 대해서는 기존 조사구에서의 실업률은 2.55%, 추가 조사구에서의 실업률은 1.97%로써 기존 조사구에서의 실업률이 높게 나타난다.

6월 자료에서 제천시의 기존 조사구의 실업률은 3.03%, 추가 조사구의 실업률은 2.50%이고 충주시의 기존 조사구의 실업률은 2.27%, 추가 조사구의 실업률은 1.23%로 나타나며, 전체 지역들에 대해서는 기존 조사구의 실업률은 2.47%, 추가 조사구의 실업률은 1.81%로써 5월 자료와 마찬가지로 기존 조사구에 대한 실업률이 높게 나타난다.

<표4.4> 충청북도 5월 시군별 자료요약

구분		경제활동인구			취업자			실업자			비경제활동인구		
		기존	추가	합계	기존	추가	합계	기존	추가	합계	기존	추가	합계
청주시	남	375	0	375	356	0	356	19	0	19	192	0	192
	여	293	0	293	286	0	286	7	0	7	339	0	339
	계	668	0	668	642	0	642	26	0	26	531	0	531
제천시	남	63	145	208	62	142	204	1	3	4	21	72	93
	여	46	96	142	45	93	138	1	3	4	55	143	198
	계	109	241	350	107	235	342	2	6	8	76	215	291
충주시	남	209	102	311	203	100	303	6	2	8	67	68	135
	여	185	58	243	182	58	240	3	0	3	131	86	217
	계	394	160	554	385	158	543	9	2	11	198	154	352
보은군	남	40	150	190	39	145	184	1	5	6	4	50	54
	여	29	123	152	29	120	149	0	3	3	28	100	128
	계	69	273	342	68	265	333	1	8	9	32	150	182
옥천군	남	79	139	218	76	134	210	3	5	8	15	51	66
	여	69	108	177	69	106	175	0	2	2	34	102	136
	계	148	247	395	145	240	385	3	7	10	49	153	202
영동군	남	46	138	184	46	136	182	0	2	2	27	36	63
	여	57	115	172	56	115	171	1	0	1	33	85	118
	계	103	253	356	102	251	353	1	2	3	60	121	181
괴산군	남	91	100	191	89	97	186	2	3	5	21	22	43
	여	84	87	171	83	87	170	1	0	1	54	63	117
	계	175	187	362	172	184	356	3	3	6	75	85	160
음성군	남	128	99	227	125	98	223	3	1	4	43	38	81
	여	92	75	167	90	73	163	2	2	4	72	68	140
	계	220	174	394	215	171	386	5	3	8	115	106	221
청원군	남	91	179	270	90	172	262	1	7	8	35	64	99
	여	79	138	217	77	137	214	2	1	3	84	123	207
	계	170	317	487	167	309	476	3	8	11	119	187	306
진천군	남	25	187	212	23	184	207	2	3	5	9	61	70
	여	20	137	157	19	136	155	1	1	2	15	137	152
	계	45	324	369	42	320	362	3	4	7	24	198	222
단양군	남	53	154	207	53	149	202	0	5	5	6	53	59
	여	41	106	147	41	106	147	0	0	0	23	108	131
	계	94	260	354	94	255	349	0	5	5	29	161	190
합계	남	1200	1393	2593	1162	1357	2519	38	36	74	440	515	955
	여	995	1043	2038	977	1031	2008	18	12	30	868	1015	1883
	계	2195	2436	4631	2139	2388	4527	56	48	104	1308	1530	2838

<표4.5> 충청북도 6월 시군별 자료요약

구분		경제활동인구			취업자			실업자			비경제활동인구		
		기존	추가	합계	기존	추가	합계	기존	추가	합계	기존	추가	합계
청주시	남	390	0	390	372	0	372	18	0	18	185	0	185
	여	288	0	288	278	0	278	10	0	10	350	0	350
	계	678	0	678	650	0	650	28	0	28	535	0	535
제천시	남	59	142	201	58	139	197	1	3	4	22	71	93
	여	40	98	138	38	95	133	2	3	5	57	136	193
	계	99	240	339	96	234	330	3	6	9	79	207	286
충주시	남	216	104	320	209	102	311	7	2	9	69	37	106
	여	181	58	239	179	58	237	2	0	2	137	86	223
	계	397	162	559	388	160	548	9	2	11	206	123	329
보은군	남	39	151	190	39	147	186	0	4	4	6	49	55
	여	29	125	154	29	123	152	0	2	2	28	102	130
	계	68	276	344	68	270	338	0	6	6	34	151	185
옥천군	남	78	145	223	77	141	218	1	4	5	15	44	59
	여	69	108	177	69	106	175	0	2	2	34	109	143
	계	147	253	400	146	247	393	1	6	7	49	153	202
영동군	남	48	142	190	48	140	188	0	2	2	27	36	63
	여	58	124	182	57	123	180	1	1	2	33	82	115
	계	106	266	372	105	263	368	1	3	4	60	118	178
괴산군	남	91	103	194	90	98	188	1	5	6	21	18	39
	여	84	91	175	84	90	174	0	1	1	54	61	115
	계	175	194	369	174	188	362	1	6	7	75	79	154
음성군	남	128	100	228	122	99	221	6	1	7	41	40	81
	여	90	75	165	89	74	163	1	1	2	76	67	143
	계	218	175	393	211	173	384	7	2	9	117	107	224
청원군	남	86	185	271	85	178	263	1	7	8	38	54	92
	여	76	141	217	75	140	215	1	1	2	90	119	209
	계	162	326	488	160	318	478	2	8	10	128	173	301
진천군	남	25	189	214	24	185	209	1	4	5	9	59	68
	여	19	137	156	18	136	154	1	1	2	16	138	154
	계	44	326	370	42	321	363	2	5	7	25	197	222
단양군	남	52	150	202	52	149	201	0	1	1	6	57	63
	여	43	112	155	43	112	155	0	0	0	21	105	126
	계	95	262	357	95	261	356	0	1	1	27	162	189
합계	남자	1212	1411	2623	1176	1378	2554	36	33	69	439	465	904
	여자	977	1069	2046	959	1057	2016	18	12	30	896	1005	1901
	계	2189	2480	4669	2135	2435	4570	54	45	99	1335	1470	2805

제 5장 시군구 실업통계 작성

5.1 개 요

통계청에서는 취업, 실업 등과 같은 경제적 특성을 조사하여 국가의 고용정책 입안과 평가에 필요한 기초자료를 수집할 목적으로 매월 3만 표본 가구 내에 거주하는 만 15세 이상인 사람들을 대상으로 경제활동인구조사를 실시하고 있다. 매월 15일을 포함하는 주 중에 표본 가구 내에 거주하는 사람들의 취업, 실업 및 비경제활동인구 관련 사항을 방문면접이나 컴퓨터면접 방식으로 조사한 후 조사된 자료를 직접추정방법에 의해 추정 계산하여 매익월 말경에 대영역인 7개 광역시와 9개 도에 대해 조사 결과를 발표하고 있다.

1995년부터 시작된 지방자치제도와 1997년 IMF 사태에 기인하여 최근에는 실업 관련 통계뿐만 아니라 다양한 분야의 통계 작성도 시군구의 소지역 단위까지 작성해야 한다는 인식이 높아지고는 있지만 현재와 같은 대영역 표본설계를 기반으로 하는 통계 작성 방법으로는 신뢰성 있는 소지역 통계 생산은 불가능한 실정이다. 왜냐하면 소지역 통계의 작성은 단순히 추정단계에서 추정량의 선택을 통해서 해결될 수 있는 것이 아니라 통계조사의 계획, 표본설계, 추정 등 통계조사 전체 과정을 종합적으로 고려할 때 가능한 일이기 때문이다.

통계청의 경제활동인구조사는 대영역인 광역시와 도별 통계 작성을 목적으로 표본설계 되었기 때문에 소지역인 시군구 단위는 표본설계에 반영된 관심영역이 아니다. 따라서 현재 활용되고 있는 대영역 기반의 표본설계를 이용하여 소지역 통계를 직접 생산할 경우 시군구 지역에 배정된 표본조사구 수가 불균형적이고 특정 조사구에 대해서는 너무 작기 때문에 신뢰할만

한 소지역 단위의 통계생산은 어렵게 된다. 우리의 목적은 대영역 기반 표본설계의 구조하에서 직접 생산된 시군구 단위의 직접추정값들을 소지역 추정에 이용되는 명시적인 모형을 통해 보정하여 어느 정도 신뢰성을 확보하는 일이다. 이를 위하여 사전에 소지역에 대한 층화, 표본 배정, 집락화의 수준에 대한 검토 후 직접추정값을 보정할 수 있는 보조정보를 고려하였다.

이 보고서에서는 현재 통계청에서 실시하고 있는 대영역 기반 경제활동인구조사의 표본설계 하에서 직접적으로 생산된 시군구 단위의 소지역에 대한 직접추정값들을 설계 기반 추정법 및 모형 기반 추정법을 통해 보정하여 시군구의 실업자 수를 추론한다. 표본 조사구 수가 불균형으로 배정된 상태에서 추정된 소지역의 월별 직접추정값들에 대해 센서스 및 행정보고자료를 통해 선택된 보조정보를 이용하여 직접추정값들에 대한 보정을 시도하였다. 설계 기반 소지역 추정법으로 합성추정법과 복합추정법이 고려되었다. 모형 기반 소지역 추정법으로는 Multi-level 모형을 이용한 계층적 베이즈 추정법과 시계열 및 횡단면 자료를 포함하는 시계열 모형을 이용한 계층적 베이즈 추정법이 고려되었다.

"Borrow Strength"를 적용하기 위하여 경제활동인구조사 자료를 시(구) 및 군 지역으로 크게 2개 그룹으로 구분하고, 각 그룹 내에서는 유사성질 범주를 성별(남, 여)-연령대별(15-34세, 35세 이상)의 4개의 범주로 구분하였다. 각 그룹에서는 성별에 따른 경제활동참가율을 그리고 각 셀에 대해서 실업률을 산출하여 해당 시군구 실업자 수 추정을 위한 보조정보로 활용하였다. 한편, 해당 셀의 경제활동인구 수를 산출하기 위하여 2000년 주민등록인구 수로부터 추계된 시군구의 성별(남, 여)-연령대별(15-34세, 35세 이상) 인구 수 자료 및 1995년 센서스 자료를 활용하였다.

5.2 소지역 추정법

5.2.1 직접추정법

통계청에서는 매달 약 30,000 표본 가구에 대해 경제활동인구조사를 실시하고 있다. 국가의 고용현황 및 경제 정책 등은 이러한 월별 조사 결과를 토대로 작성된다. 통계청의 경제활동인구조사에 대한 관심영역은 7 개 광역시와 9 개 도 단위를 포함한 16 개의 대영역이다.

이러한 지역들에 대한 실업자 총계를 추정하기 위한 직접추정량은 다음과 같은 총계 추정공식이 이용된다.

$$\hat{Y}_{i\cdot} = \sum_{s=1}^{2} {}_s\hat{Y}_{i\cdot} \quad , \; i=1,2,\cdots,I \; ; s=1,2 \; ; h=1,2,\cdots,n_i$$

$$= \sum_{s=1}^{2}\sum_{h=1}^{n_i} {}_s\hat{Y}_{ih}$$

$$= \sum_{s=1}^{2}\sum_{h=1}^{n_i} {}_sM_i \; {}_sY_{ih} \qquad (5.1)$$

여기에서 s : 성별(남-여)을 나타내는 첨자

n_i : 경제활동인구조사에서 i번째 지역의 표본조사구 수

${}_sY_{ih}$: 각 성별에 대해서 i번째 지역의 표본 조사구에서 조사한 실업자 수를 나타낸다.

승수 ${}_sM_i = {}_s\hat{X}_{i\cdot} / {}_sX_{i\cdot}$ 은 $\hat{Y}_{i\cdot}$ 이 불편추정량이 되도록 산정한다.

여기에서 ${}_s\hat{X}_{i\cdot}$: i번째 지역에 대한 15세 이상의 상주 추계인구

${}_sX_{i\cdot}$: 경제활동인구조사에서 조사된 15세 이상의 조사인구

직접추정량 $\hat{Y}_{i\cdot}$의 분산은 다음 (5.2)식과 같이 주어진다.

$$Var(\hat{Y}_{i\cdot}) = \sum_{s=1}^{2} Var({}_s\hat{Y}_{i\cdot}) + 2Cov({}_1\hat{Y}_{i\cdot}, {}_2\hat{Y}_{i\cdot}), \quad i=1,2,\cdots,I$$

$$= \sum_{s=1}^{2} {}_sM_i^2 Var(\sum_{h=1}^{n_i} {}_sY_{ih}) + 2\,{}_1M_i\,{}_2M_i Cov(\sum_{h=1}^{n_i} {}_1Y_{ih}, \sum_{h=1}^{n_i} {}_2Y_{ih}). \quad (5.2)$$

$\hat{Y}_{i\cdot}$의 분산에 대한 추정은 다음 (5.3)식의 추정공식을 이용하여 계산된다.

$$\hat{Var}(\hat{Y}_{i\cdot}) = \sum_{s=1}^{2} {}_sM_i^2(\zeta_i \sum_{h=1}^{n_i} {}_sU_{ih}^2) + 2\,{}_1M_i\,{}_2M_i(\zeta_i \sum_{h=1}^{n_i} {}_1U_{ih}\,{}_2U_{ih}), \quad (5.3)$$

여기에서 ${}_sU_{ih} = d_sY_{ih} - {}_s\rho_i \cdot d_sX_{ih}$, $d_sY_{ih} = {}_sY_{ih} - {}_sY_{i,h+1}$,

$d_sX_{ih} = {}_sX_{ih} - {}_sX_{i,h+1}$, ${}_s\rho_i = {}_sY_{i\cdot} / {}_sX_{i\cdot}$,

$\zeta_i = [1 - n_i/(10N_i)]n_i/[2(n_i-1)]$ 이고, N_i는 소지역 i에 대한 모집단의 조사구 수를 나타낸다.

경제활동인구조사에서 대영역에 포함된 소지역들은 표본설계에 반영된 관심영역이 아니다. 따라서 대영역 표본설계에 기반을 둔 표본 조사로부터 소지역들에 대한 관심 통계량들을 추정한다면 소지역에 할당된 표본 조사구 수가 충분하지 않기 때문에 신뢰할 만한 결과를 얻을 수 없다. 이러한 관점에서 볼때 경제활동인구조사에서 소지역에 대한 직접추정값은 목표 정도를 만족할 수 없게 된다.

이 보고서의 목적은 경제활동인구조사에서 추정된 소지역에 대한 직접추정값을 설계 기반 추정량 또는 모형 기반 추정량을 통해 보정하여 어느 정도의 신뢰성을 확보하는 일이다. 이를 위하여 층화, 표본 할당, 집락의 수준, 소지역의 직접추정값을 보정하기 위한 보조정보 등이 검토되었다.

5.2.2 합성 추정법

대영역 표본설계에 기반을 둔 통계청의 직접추정량은 각 소지역에 할당된 표본 조사구의 수가 충분하지 않기 때문에 소지역 실업통계의 정확도를 제공하지는 못한다. 합성추정량 $\hat{Y}_{i\cdot}^{S}$는 인근 유사지역의 정보를 소지역 추정에 이용하는 간접적인 설계 기반 추정량이다.

대영역을 I개의 시군구 단위의 소지역들로 분할하고, 대영역을 특성 기준에 따라 유사성을 갖는 J개의 성별-연령대별 범주들로 구분할 때 i번째 소지역의 합성추정량 $\hat{Y}_{i\cdot}^{S}$는 다음과 같이 주어질 수 있다.

$$\hat{Y}_{i\cdot}^{S} = \sum_{j=1}^{J} \eta_{ij} \Psi_{\cdot j\cdot}^{a} \quad , \quad i=1,2,\cdots,I \ , \tag{5.4}$$

(5.4)식에서 가중치 $\eta_{ij} = (\xi_{ijt}^{C} / \xi_{ijt}^{R})\, \xi_{ij}\, {}_{s}x_{j}$ 는 i번째 소지역에서 j번째 범주에 대한 경제활동 추정인구를 나타낸다. 여기에서 ξ_{ijt}^{C}는 t번째 해의 상주추정인구, ξ_{ijt}^{R}는 같은 해의 주민등록인구, ${}_{s}x_{j}$는 경제활동인구조사에서 각 성별에 대한 j번째 범주의 경제활동 참가율을 나타낸다.

$\Psi_{\cdot j\cdot}^{a} = \hat{Y}_{\cdot j\cdot} / \sum_{i=1}^{I} \phi_{ij}$ $(j=1,2,\cdots,J)$는 경제활동인구조사에서 추정된 j번째 범주에 대한 실업률을 나타낸다. 여기에서 $\hat{Y}_{\cdot j\cdot} = \sum_{i=1}^{I} \sum_{h=1}^{n_i} M_j Y_{ijh}$는 j번째 범주에 대한 직접추정량, $M_j = \hat{X}_{\cdot j\cdot} / X_{\cdot j\cdot}$ 는 j번째 범주에 대한 승수 값, ϕ_{ij}는 i번째 소지역에서 j번째 범주에 대한 경제활동인구를 나타낸다.

소지역 i에서 j번째 범주에 대한 경제활동인구 η_{ij}를 상수로 가정한다면 합성추정량 $\hat{Y}_{i\cdot}^{S}$의 분산은 다음 (5.5)식과 같이 주어질 수 있다.

$$Var(\hat{Y}_{i\cdot}^{S}) = \sum_{j=1}^{J} \eta_{ij}^2 Var(\Psi_{j\cdot}^{a}) + 2\sum_{j<l} \eta_{ij}\eta_{il} Cov(\Psi_{j\cdot}^{a}, \Psi_{l\cdot}^{a}), i=1,2,\cdots,I \quad (5.5)$$

합성추정량 $\hat{Y}_{i\cdot}^{S}$의 추정분산은 다음 (5.6)식으로부터 계산될 수 있다.

$$\begin{aligned} \hat{Var}(\hat{Y}_{i\cdot}^{S}) &= \sum_{j=1}^{J} \eta_{ij}^2 \left(\frac{1}{\sum_{i=1}^{I}\phi_{ij}}\right)^2 \hat{Var}(\hat{Y}_{\cdot j\cdot}) \\ &\quad + 2\sum_{j<l} \eta_{ij}\eta_{il} \left(\frac{1}{\sum_{i=1}^{I}\phi_{ij}}\right)\left(\frac{1}{\sum_{i=1}^{I}\phi_{il}}\right) \hat{Cov}(\hat{Y}_{\cdot j\cdot}, \hat{Y}_{\cdot l\cdot}) \\ &= \sum_{j=1}^{J} \eta_{ij}^2 \left(\frac{1}{\sum_{i=1}^{I}\phi_{ij}}\right)^2 \left(M_j^2 \zeta_j \sum_{i=1}^{I}\sum_{k=1}^{n_{ij}} U_{ijk}^2\right) \\ &\quad + 2\sum_{j<l} \eta_{ij}\eta_{il} \left(\frac{1}{\sum_{i=1}^{I}\phi_{ij}}\right)\left(\frac{1}{\sum_{i=1}^{I}\phi_{il}}\right)\left(M_j M_l \zeta_j \sum_{i=1}^{I}\sum_{k=1}^{n_{ij}} U_{ijk}U_{ilk}\right), \quad (5.6) \end{aligned}$$

여기에서 $U_{ijk} = d_j Y_{ijk} - \rho_j \cdot d_j X_{ijk}$, $d_j Y_{ijk} = Y_{ijk} - Y_{ij,k+1}$,

$d_j X_{ijk} = X_{ijk} - X_{ij,k+1}$, $\rho_j = Y_{\cdot j\cdot} / X_{\cdot j\cdot}$,

$\zeta_j = [1 - n_j/(10N_j)]\, n_j/[2(n_j - 1)]$. n_j는 경제활동인구조사에서 j번째 범주에 대한 표본 조사구 수를 나타내며, N_j는 j범주에 대한 모집단의 조사구 수를 나타낸다.

5.2.3 복합 추정법

대영역을 기반으로한 기존의 직접추정값들은 해당 소지역에 할당된 표본수가 충분하지 않기 때문에 추정값들의 신뢰성을 확보할 수 없다. 또한 인근지역의 유사정보를 이용하여 추정되는 합성추정값들은 잠재적인 편향 가능성이 항상 내재되어 있다. 따라서 소지역에 배정된 표본 수가 적을 경우, 표본조사만을 이용하여 추정되는 직접추정량의 불안정성과 합성추정량의 편향

가능성을 보완하기 위해 직접추정량과 합성추정량의 가중평균을 이용한 다음과 같은 복합추정량이 고려될 수 있다.

$$\hat{Y}_{i\cdot}^{C} = \omega_i \hat{Y}_{i\cdot} + (1-\omega_i) \hat{Y}_{i\cdot}^{S}, \quad i=1,2,\cdots,I, \qquad (5.7)$$

여기에서 $\hat{Y}_{i\cdot}$은 i번째 소지역의 실업자 총계에 대한 직접추정량, $\hat{Y}_{i\cdot}^{S}$는 인근 유사 지역의 정보를 이용하여 추정되는 합성추정량을 나타내며 가중치 ω_i는 0과 1사이의 값을 취한다.

최적 가중치 $\omega_{i(opt)}$는 $Cov(\hat{Y}_{i\cdot}, \hat{Y}_{i\cdot}^{S})=0$를 가정하여 $MSE(\hat{Y}_{i\cdot}^{C})$를 최소화시키는 가중값으로 결정되며 다음 (5.8)식과 같이 주어진다.

$$\omega_{i(opt)} = \frac{MSE(\hat{Y}_{i\cdot}^{S})}{MSE(\hat{Y}_{i\cdot}^{S}) + Var(\hat{Y}_{i\cdot})}, \quad i=1,2,\cdots,I \qquad (5.8)$$

여기에서 최적 가중치 $\omega_{i(opt)}$은 $Var(\hat{Y}_{i\cdot})$과 $MSE(\hat{Y}_{i\cdot}^{S})$의 함수로 표현되며 추정되어야 할 값들이다.

합성추정량 $\hat{Y}_{i\cdot}^{S}$의 분산은 (5.6)식을 이용하여 추정할 수 있으나 $MSE(\hat{Y}_{i\cdot}^{S})$의 정확한 추정공식은 아직까지는 유도된 바가 없다. 근사값으로 다음과 같은 $MSE(\hat{Y}_{i\cdot}^{S})$의 근사적인 불편추정량이 이용되기도 한다.

$$mse(\hat{Y}_{i\cdot}^{S}) \cong (\hat{Y}_{i\cdot}^{S} - \hat{Y}_{i\cdot})^2 - \hat{V}ar(\hat{Y}_{i\cdot}), \quad i=1,2,\cdots,I \qquad (5.9)$$

그러나 다단계 표본에서는 다수의 소지역들이 하나 또는 둘 정도의 일차추출단위 만을 포함하고 있기 때문에 추정분산 $\hat{V}ar(\hat{Y}_{i\cdot})$은 큰 변동을 나타내서 조사구 수가 적은 소지역에서는 $mse(\hat{Y}_{i\cdot}^{S})$의 값이 음수를 가질 가능성이 있다.

이러한 관점에서 하나의 대안으로 $mse(\hat{Y}_{i\cdot}^{S})$를 $\hat{V}ar(\hat{Y}_{i\cdot}^{S})$로 대체하여 다음과 같은 추정 가중치를 이용하였다.

$$\hat{\omega}_{i(opt)} = \frac{\hat{V}ar(\hat{Y}_{i\cdot}^{S})}{\hat{V}ar(\hat{Y}_{i\cdot}^{S}) + \hat{V}ar(\hat{Y}_{i\cdot})} \quad , \quad i=1,2,\cdots,I \qquad (5.10)$$

따라서 i번째 소지역에 대한 복합추정량 $\hat{Y}_{i\cdot}^{C}$는 다음 (5.11)식을 이용하여 추정될 수 있다.

$$\hat{Y}_{i\cdot}^{C} = \hat{\omega}_{i(opt)}\ \hat{Y}_{i\cdot} + (1-\hat{\omega}_{i(opt)})\ \hat{Y}_{i\cdot}^{S} \quad , \quad i=1,2,\cdots,I \qquad (5.11)$$

직접추정량과 합성추정량의 공분산 $Cov(\hat{Y}_{i\cdot}, \hat{Y}_{i\cdot}^{S})=0$의 가정 하에 복합추정량의 추정분산은 다음 (5.12)식으로부터 계산될 수 있다.

$$\hat{V}ar(\hat{Y}_{i\cdot}^{C}) = \hat{\omega}_{i(opt)}^{2}\ \hat{V}ar(\hat{Y}_{i\cdot}) + (1-\hat{\omega}_{i(opt)})^{2}\ \hat{V}ar(\hat{Y}_{i\cdot}^{S}) \qquad (5.12)$$

5.2.4 Multi-level 모형을 이용한 계층적 베이즈 추정법

대영역은 I개의 소지역을 포함하고 있고, 이러한 대영역에 대해 경제활동인구조사 자료에서 K개의 월별 조사 자료들이 임의로 선택되었다고 하자. 이때 소지역들 간의 변동과 소지역들 내에서의 변동을 하나의 모형으로 통합한 다음과 같은 Multi-level 모형을 고려할 수 있다.

$$y_{ik} = x_{ik}^{T}\beta_i + e_{ik} \quad , \quad \beta_i = Z_i\gamma + \nu_i \quad , \quad i=1,2,\cdots,I\ ;\ k=1,2,\cdots,K \qquad (5.13)$$

Multi-level 모형 (5.13)에서 y_{ik}는 i번째 소지역에서 k번째 월에 실시된

경제활동인구조사의 직접추정값을 나타내며, 이 값들은 경제활동인구조사, 센서스 및 행정보고 자료로부터 선택된 보조변수

$x_{ik}=(x_{i1k}, x_{i2k}, \cdots, x_{ipk})^T$를 이용하여 모형 (5.13)을 통해 보정된다. β_i는 $p\times 1$ 회귀 계수 벡터, Z_i는 $p\times q$ 설계 행렬, γ는 $q\times 1$ 고정 계수 벡터, $\nu_i=(\nu_{i1}, \nu_{i2}, \cdots, \nu_{ip})^T$는 i번째 소지역에 대한 $p\times 1$ 랜덤효과 벡터를 나타낸다.

ν_i는 결합분포 $\nu_i \overset{ind}{\sim} N_p(0, \Phi)$를 가지며, 여기에서 분산 공분산 행렬 Φ는 미지인 값이다. 오차항 e_{ik}는 평균 $E(e_{ik})=0$와 분산 $Var(e_{ik})=\sigma_i^2$를 갖는 서로 독립인 확률변수들이고, ν_i와 e_{ik}는 서로 독립임을 가정한다.

소지역에 대한 계층적 베이즈 추정값과 추정값의 사후분산을 계산하기 위해 계층적 베이즈 Multi-level 모형 구조를 다음과 같이 설정할 수 있다.

$$(\mathrm{i})\ [y_{ik}|\ \beta_i, \sigma_i^2] \overset{ind}{\sim} N(x_{ik}\beta_i, \sigma_i^2),\ i=1,2,\cdots,I;\ k=1,2,\cdots,K \quad (5.14)$$

$$(\mathrm{ii})\ [\beta_i|\gamma, \Phi] \overset{ind}{\sim} N_p(Z_i\gamma, \Phi),\ i=1,2,\cdots,I \quad (5.15)$$

(iii) 주변 확률사전분포는 다음과 같이 가정한다:

$$\gamma \sim N_q(0, D),\quad \tau_i \overset{ind}{\sim} G(a_i, b_i),\quad \Omega \sim W_p(\alpha, R).$$

여기에서 $\tau_i=\sigma_i^{-2}$, $\Omega=\Phi^{-1}$을 나타내며, D, a_i, b_i, α, R은 주어지는 값이다. $G(a_i, b_i)$는 밀도함수 $f(x)=[b_i^{a_i}/\Gamma(a_i)]x^{a_i-1}e^{-b_ix}$ ($a_i>0$, $b_i>0$, $x\geq 0$)를 갖는 감마분포이다.

i번째 소지역에서 k번째 달에 대한 사후추정치 $\mu_{ik} = \boldsymbol{x}_{ik}^T \boldsymbol{\beta}_i$는 주어진 $y=(\{y_{ik}\}, i=1,2,\cdots,I; k=1,2,\cdots,K)$에 대한 $\boldsymbol{\beta}_i$의 사후분포로부터 생성된 깁스 표본을 이용하여 계산될 수 있다. 이러한 과정은 모수들에 대한 완전한 조건부 사후분포로부터 수행될 수 있다.

앞서 언급한 계층적 베이즈 Multi-level 모형 구조하에서 모수들에 대한 깁스 반복 표본을 생성하기 위해 필요한 조건부 사후분포들은 다음과 같이 주어질 수 있다.

(i) $[\boldsymbol{\beta}_i | y, \gamma, \Omega, \tau] \overset{\text{ind}}{\sim} N_p\Big((\tau_i \sum_k \boldsymbol{x}_{ik} \boldsymbol{x}_{ik}^T + \Omega)^{-1}(\tau_i \sum_k y_{ik} \boldsymbol{x}_{ik} + \Omega Z_i \gamma), (\tau_i \sum_k \boldsymbol{x}_{ik} \boldsymbol{x}_{ik}^T + \Omega)^{-1}\Big)$,

(ii) $[\gamma | y, \beta, \Omega, \tau] \sim N_q\Big((\sum_i Z_i^T \Omega Z_i + D^{-1})(\sum_i Z_i^T \Omega \boldsymbol{\beta}_i), (\sum_i Z_i^T \Omega Z_i + D^{-1})^{-1}\Big)$,

(iii) $[\Omega | y, \beta, \gamma, \tau] \sim W_p\Big(a+I, R+\frac{1}{2}\sum_i (\boldsymbol{\beta}_i - Z_i\gamma)(\boldsymbol{\beta}_i - Z_i\gamma)^T\Big)$,

(iv) $[\tau_i | y, \beta, \gamma, \Omega] \overset{\text{ind}}{\sim} G\Big(a_i + \frac{K}{2}, b_i + \frac{1}{2}\sum_k (y_{ik} - \boldsymbol{x}_{ik}^T \boldsymbol{\beta}_i)^2\Big)$.

위의 조건부 사후분포에서 $\gamma^{(0)}$, $\Omega^{(0)}$, $\tau_i^{(0)}$를 추정된 초기값으로 하여 과정(i)에서 $\boldsymbol{\beta}_i^{(1)}$을 생성하고 과정(ii)에서 $\gamma^{(1)}$을, 과정(iii)과 (iv)에서 $\Omega^{(1)}$와 $\tau_i^{(1)}$을 각각 생성하는 깁스 샘플링을 반복적으로 수행할 수 있다. 이때 안정적인 깁스 샘플링 자료를 얻기 위해 "Burn-in" 주기 이후의 M개의 깁스 표본 $\{\boldsymbol{\beta}_i^{(m)}, \gamma^{(m)}, \Omega^{(m)}, \tau_i^{(m)}; m=1,2,\cdots,M\}$을 $(\boldsymbol{\beta}_i, \gamma, \Omega, \tau_i)$의 결합사후분포로부터 추출한 반복표본으로 간주한다. 여기에서 $\boldsymbol{\beta}_i$의 사후추정값은

생성된 M개의 반복표본 $\{ \beta_i^{(m)} ; m=1,2,\cdots,M\}$을 이용하여 계산될 수 있다.

μ_{ik}의 사후평균 추정값과 추정값들의 분산은 MCMC(Markov Chain Monte Carlo)기법을 이용하여 추정할 수 있으며 본연구에서는 WinBUGS 소프트웨어(Spiegelhalter et al. 2000)를 통해 계산하고자 한다.

5.2.5 시계열 및 횡단면 모형을 이용한 계층적 베이즈 추정법

대영역 내에 I개의 소지역이 존재하고 월별 경제활동인구조사 자료가 연속적으로 선택되었다고 가정할 때, 소지역 랜덤효과 ν_i와 $AR(1)$ 과정 u_{it}를 갖는 다음과 같은 선형혼합모형을 고려할 수 있다.

$$y_{it} = \boldsymbol{x}_{it}^T \boldsymbol{\beta} + \nu_i + u_{it} \ , \quad i=1,2,\cdots,I \ ; \ t=1,2,\cdots,T$$

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it} \ , \quad |\rho| < 1 \ , \tag{5.16}$$

모형 (5.16)에서 y_{it}는 i번째 소지역에서 t번째 달에 실시한 경제활동인구조사의 직접추정값을 나타내며, 이 값들은 경제활동인구조사, 센서스, 행정보고자료 등을 이용하여 선택된 횡단면 변량 $\boldsymbol{x}_{it}=(x_{i1t}, x_{i2t}, \cdots, x_{ipt})^T$에 의해 모형 (5.16)을 통해 보정된다. 여기에서 $\boldsymbol{\beta}$는 $p\times 1$ 회귀계수 백터, ν_i는 $\nu_i \overset{iid}{\sim} N(0, \sigma_i^2)$ 의 분포를 따르는 소지역 랜덤효과, u_{it}는 $AR(1)$ 과정, 오차항 $\varepsilon_{it} \overset{iid}{\sim} N(0, \sigma^2)$이며 $\{\nu_i\}$와 $\{\varepsilon_{it}\}$는 서로 독립임을 가정하자.

모형 (5.16)을 계층적 베이즈 모형 구조로 정리하면 다음과 같아질 것이다. 여기에서 $r_i = \sigma_i^{-2}$, $r = \sigma^{-2}$을 나타낸다.

$$(\mathrm{i})\ [y_{it} \mid \nu_i, \beta, r_i, r, \rho] \overset{\text{ind}}{\sim} N\left(\boldsymbol{x}_{it}^{T}\beta + \nu_i,\ \frac{r^{-1}}{1-\rho^2} \right), \tag{5.17}$$

$$(\mathrm{ii})\ [\nu_i \mid \beta, r_i, r, \rho] \overset{\text{ind}}{\sim} N(0, r_i^{-1}), \tag{5.18}$$

(iii) 주변 사전확률분포는 다음과 같이 가정한다:

$\beta \sim Uniform(R^p)$, $r \sim G\left(\frac{a}{2}, \frac{b}{2}\right)$, $r_i \sim G\left(\frac{c_i}{2}, \frac{d_i}{2}\right)$, $\rho \sim f(\rho)$, 여기에서 ρ의 사전확률분포는 적절한 임의의 분포로 가정되며 $G(a, b)$는 확률밀도함수 $f(x) = [b^a/\Gamma(a)]x^{a-1}e^{-bx}$ $(a>0, b>0, x \geq 0)$를 갖는 감마분포이다.

i번째 소지역에서 t번째 달에 실시한 경제활동인구조사 자료에 대한 $\mu_{it} = \boldsymbol{x}_{it}^{T}\beta + \nu_i$ 의 사후추정값들을 계산하기 위해 깁스 반복 표본을 이용할 수 있다. 주어진 y에 대해서 μ_{it}의 사후평균 추정값들과 추정값들의 분산은 모수들에 대한 조건부 확률분포로부터 생성된 반복표본을 이용하여 MCMC 적분기법을 통해 계산할 수 있다.

깁스 반복표본을 생성하기 위해 필요한 조건부 확률분포는 다음과 같이 주어진다.

$$(\mathrm{i})\ [\beta \mid \nu_i, r_i, r, \rho, y] \sim N_p\Big(\Big(\sum_i \sum_t \boldsymbol{x}_{it}\ \boldsymbol{x}_{it}^{T}\Big)^{-1} \sum_i \sum_t (y_{it}\ \boldsymbol{x}_{it} - \nu_i\ \boldsymbol{x}_{it}),$$

$$[r(1-\rho^2)]^{-1} \Big(\sum_i \sum_t \boldsymbol{x}_{it}\ \boldsymbol{x}_{it}^{T}\Big)^{-1} \Big),$$

(ii) $[\nu_i|\beta, r_i, r, \rho, y] \overset{ind}{\sim} N\Big([r(1-\rho^2)+r_i][r(1-\rho^2)]^{-1}\sum_t(y_{it}- x_{it}^T\beta),$

$[r(1-\rho^2)+r_i]^{-1}\Big)$

(iii) $[r_i|\beta, \nu_i, r, \rho, y] \overset{ind}{\sim} G\Big(\frac{1}{2}(c_i+\nu_i^2), \frac{1}{2}(d_i+1)\Big)$,

(iv) $[r|\beta, \nu_i, r_i, \rho] \sim G\Big(\frac{1}{2}\Big[a+(1-\rho^2)\sum_i\sum_t(y_{it}- x_{it}^T\beta-\nu_i)^2\Big]$,

$\frac{1}{2}(IT+b)\Big)$,

(v) $[\rho|\beta, \nu_i, r_i, r, y]=[A(\beta, \nu_i, r)]^{-1}(1-\rho^2)^{\frac{n}{2}}f(\rho)$

$\times\exp\Big\{-\frac{r(1-\rho^2)}{2}\sum_i\sum_t(y_{it}- x_{it}^T\beta-\nu_i)^2\Big\}$,

단, $A(\beta, \nu_i, r)=\int(1-\rho^2)^{\frac{n}{2}}\exp\Big\{-\frac{r(1-\rho^2)}{2}\sum_i\sum_t(y_{it}- x_{it}^T\beta-\nu_i)^2\Big\}$

$\times f(\rho)\,d\rho$.

초기값 $\beta^{(0)}$, $\nu_i^{(0)}$, $r_i^{(0)}$, $r^{(0)}$를 이용하여 위의 조건부 확률분포 (i)-(v)로부터 깁스 반복표본을 생성할 수 있다. 조건부 확률분포 (v)의 확률표본은 수치적분 $A(\beta, \nu_i, r)$을 직접적으로 계산하지 않고 채택 기각 표본추출 기법을 이용하여 반복 생성할 수 있다(Zeger and Karim, 1991). 안정적인 깁스 샘플링 자료를 얻기 위해 "Burn-in" 주기 이후의 M개의 깁스 표본 $\{\beta^{(m)}, \nu_i^{(m)}, r_i^{(m)}, r^{(m)}, \rho^{(m)}; m=1,2,\cdots,M\}$을 모수 $(\beta, \nu_i, r_i, r, \rho)$의 결합확률분포로부터 생성된 반복표본으로 간주하여 특성값에 대한 추정이 이루어진다. 주어진 y에 대해 $\mu_{it}= x_{it}^T\beta+\nu_i$의 사후평균 추정값과 추정량의 분산은 M개의 깁스 반복표본 $\{\beta^{(m)}, \nu_i^{(m)}; m=1,2,\cdots,M\}$으로부터 MCMC 적분기법을 이용하여 계산할 수 있다.

5.3 시군구 실업통계 작성 예

5.3.1 서 론

이 보고서에서는 대영역인 광주광역시 5개 구와 충청북도 내의 11개 시군구 단위의 소지역에 대한 실업통계 작성 절차를 설명하기로 한다. 2001년 5월과 6월의 경제활동인구조사 및 상주추정인구를 이용하여 추정된 광주광역시 및 충청북도 내의 시군구 단위의 소지역에 대한 경제활동인구 및 각 시군구 단위에 대한 표본 조사구 수가 다음 <표5.1>과 <표5.2>에 주어졌다.

<표5.1>에서 2001년 5월의 광주광역시의 경제활동인구는 593,329명이며 이중 남자는 57.8%인 342,333명이고 여자는 250,996명으로 42.2%를 차지하였으며 각 구별 경제활동인구 남자 구성비는 57.3%에서 58.3%사이에 있으며 평균적으로 58%에 집중되어 있으며 구별간의 차이가 거의 없는 것으로 나타났다. 여자의 경우에는 구별 구성비가 42%근처에 집중된 특성을 보이고 있다.

2001년 6월의 경제활동인구는 591,359명으로 5월에 비해서 1,970명이 감소했으나 여자는 오히려 250,996명으로 18,006명이 증가하였다. 6월의 남자의 각구별 구성비는 54.1%에서 55.1%사이에 분포되어있으며 55%근처에 집중된 특성을보이고 있으며 여자의 경우도 44.9%에서 45.9%사이에 분포되어 있고 각구별로는 구성비에서 차이가 없는 것으로 나타나 있다.

<표 5.1> 광주광역시의 경제활동인구와 표본조사구 수

( ): 구성비

시군구	경제활동인구(5월)			경제활동인구(6월)			표본 조사구수
	남자	여자	합계	남자	여자	합계	
동구	33,094	23,698	56,792	31,061	25,305	56,366	10
	(0.583)	(0.417)		(0.551)	(0.449)		
서구	71,976	53,205	125,181	68,069	57,299	125,368	15
	(0.575)	(0.425)		(0.543)	(0.457)		
남구	59,940	44,665	104,605	56,308	47,739	104,047	16
	(0.573)	(0.427)		(0.541)	(0.459)		
북구	120,225	87,729	207,954	113,106	93,923	207,029	34
	(0.578)	(0.422)		(0.546)	(0.454)		
광산구	57,098	41,699	98,797	53,813	44,736	98,549	12
	(0.578)	(0.422)		(0.546)	(0.454)		
합계	342,333	250,996	593,329	322,357	269,002	591,359	87
	(0.578)	(0.422)		(0.545)	(0.455)		

<표5.2>에서 충청북도의 2001년 5월의 경제활동인구는 580,949명이고 이중에서 남자는 344,234명으로 59.3%이며 여자는 236,715명으로 40.7%이다. 청주시의 경제활동인구는 260,917명으로 도전체의 44.9%를 차지하고 있으며 남자의 구성비는 57.6%이고 여자의 구성비는 42.4%로 타 시군의 여자 구성비가 38.5%에서 40.5%인 것에 비하면 여자의 경제활동참여가 활발한 것으로 보인다.

2001년 6월에는 경제활동인구가 572,279명으로 5월에 비해서 8,670명이 감소하였지만 각 시군별 성별간의 차이는 5월과 유사하게 나타나 있다. 6월에는 여자에 비해서 남자들의 각 시군에서 구성비가 약 1%정도 전체적으로 높은 것으로 나타나 있다.

광주광역시와 충북을 비교하면 도시지역일수록 여자의 경제활동참여율이 높아지고 있으며 충북에서 괴산군과 영동군의 여자의 경제활동참여율이 38.3%와 38.4%로 가장 낮은 것으로 나타나 있다.

<표 5.2> 충청북도의 경제활동인구와 표본조사구 수

( ): 구성비

시군구		경제활동인구(5월)			경제활동인구(6월)			표본조사구수
		남자	여자	합계	남자	여자	합계	
시지역	청주시	150,206	110,711	260,917	152,574	102,599	255,173	22
		(0.576)	(0.424)		(0.598)	(0.402)		
	충주시	48,256	30,474	78,730	48,873	29,423	78,296	11
		(0.613)	(0.387)		(0.624)	(0.376)		
	제천시	30,520	19,900	50,420	30,932	19,234	50,166	4
		(0.605)	(0.395)		(0.617)	(0.383)		
군지역	보은군	7,400	4,711	12,111	7,327	4,611	11,938	2
		(0.611)	(0.389)		(0.614)	(0.386)		
	옥천군	13,063	8,576	21,639	12,949	8,413	21,362	3
		(0.604)	(0.396)		(0.606)	(0.394)		
	영동군	13,144	8,294	21,438	13,014	8,126	21,140	3
		(0.615)	(0.385)		(0.617)	(0383)		
	괴산군	7,276	4,553	11,829	7,203	4,465	11,668	5
		(0.609)	(0.391)		(0.617)	(0.383)		
	음성군	21,574	13,841	35,415	21,417	13,600	35,017	6
		(0.609)	(0.391)		(0.598)	(0.402)		
	청원군	30,232	20,466	50,698	30,046	20,135	50,181	5
		(0.596)	(0.404)		(0.598)	(0.402)		
	진천군	14,328	9,750	24,078	14,259	9,605	23,864	1
		(0.595)	(0.405)		(0.598)	(0.402)		
	단양군	8,235	5,439	13,674	8,150	5,324	13,474	2
		(0.602)	(0.398)		(0.605)	(0.395)		
합계		344,234	236,715	580,949	346,744	225,535	572,279	64
		(0.593)	(0.407)		(0.606)	(0.394)		

"Borrow Strength"를 적용하기 위해 경제활동인구 조사에서 대영역인 광주 광역시는 한 그룹으로 간주하고, 충청북도는 시지역과 군지역의 두개 그룹으로 나누고, 각각의 시군구 그룹은 유사성질을 갖는 성별(남, 여)-연령대별(15-34세, 35세이상)의 4개의 범주로 구분하였다.

경제활동인구 표본조사를 이용하여 각 셀에 대해서 추정된 경제활동인구와 실업률이 다음 <표5.3>과 <표5.4>에 주어졌다. 각 셀에 대해서 추정된 실업률은 시군구 단위의 소지역들에 대한 실업자 수를 추정하기 위한 보조정보로 활용되었다.

범주를 4개로 나눈 이유는 우리나라 전체의 2001년 5월의 남자 실업률은 3.9%이고 여자는 2.8%이며 연령별로는 15-19세의 실업률은 11.7%이고 20-29세의 실업률은 6.7%이며 30-39세의 실업률은 3.2%, 40세이상의 실업률은 3.0%이내이기 때문에 35세를 구분의 기준으로 하였다.

<표 5.3> 광주광역시의 경제활동인구와 실업률

구 분	연령	경제활동인구		실업률	
		남자	여자	남자	여자
5월	15-34	106,077	101,708	0.0720	0.0599
	35+	256,0236	149,288	0.0297	0.0292
6월	15-34	114,226	113,559	0.0717	0.0836
	35+	208,131	155,443	0.0268	0.0344

<표 5.3>에서 남자 경제활동인구가 전체적으로는 6월에 감소하였으나 34세 이하에서는 오히려 5월 106,077명에서 6월 114,226명으로 8,149명이 증

가하였으며 35세 이상의 남자구성비가 상대적으로 높게 나타났다.

5월의 실업률은 남자가 34세 이하에서 7.2%이고 35세 이상에서는 2.97%로 청년층의 실업률이 중장년의 배 이상이며 여자는 34세 이하가 5.99%, 35세 이상은 2.92%로 연령층간에는 실업률이 큰 차이를 보이고 있으나 35세 이상에서는 남자와 여자간의 차이가 심각하지 않다.

6월의 실업률은 여자는 34세 이하에서 8.36%로 많이 증가하였고 남자는 35세 이상에서 0.3%정도 낮아졌다.

<표 5.4>에서 충북지역의 경제활동인구에서 34세 이하는 남자와 여자간의 차이가 작지만 35세 이상에서는 남자가 훨씬 많은 것으로 나타났고 군지역에서는 35세 이상에서도 남자와 여자간의 경제활동인구의 차이가 크지않음을 볼 수 있다. 특히 군지역에서는 5월과 6월간의 차이가 별로 없으므로 군지역의 실업률 특성을 짐작할 수 있다.

실업률에서 5월 남자의 34세 이하가 6.17%로 높고 남자 35세 이상은 3.83%로 낮지만 같은 연령층의 여자에 비해서는 3배 이상 높은 숫자이다. 군지역에서 34세 이하의 남자 실업률은 4.23%이고 여자의 실업률은 4.4%로 비슷하며 6월에도 남자는 3.57%이고 여자는 3.49%로 5월과 같이 별다른 차이를 보이지 않고 있다. 그러나 군지역의 35세 이상은 남자가 1.39%로 여자의 0.65%의 2배 이상이고 6월에도 여자는 0.22%, 남자는 1.2%로 5배 이상의 차이를 나타내고 있다.

결론적으로 군지역에서 34세 이하는 남녀간의 실업률 차이가 크지 않은 반면에 35세 이상에서는 남자의 실업률이 여자보다 훨씬 높게 나타나고 있다.

<표 5.4> 충북 경제활동인구와 실업률

구분	그 룹	연령	경제활동인구		실업률	
			남자	여자	남자	여자
5월	시지역	15-34	63,048	57,578	0.0617	0.0503
		35+	152,951	102,481	0.0382	0.0106
	군지역	15-34	37,377	22,943	0.0423	0.0440
		35+	132,399	116,479	0.0139	0.0065
6월	시지역	15-34	65,473	56,778	0.0647	0.0510
		35+	154,438	98,367	0.0349	0.0221
	군지역	15-34	36,824	21,505	0.0357	0.0349
		35+	131,776	115,525	0.0120	0.0022

5.3.2 합성 및 복합 추정량

i번째 시군구 단위의 소지역의 실업자에 대한 직접 추정값과 추정량의 분산 추정값은 각각 식(5.1)과 (5.3)에 의해 계산된다. 실업자 총계에 대한 합성 추정값은 $\hat{Y}_{i\cdot}^{S}=\sum_{j=1}^{4}\eta_{ij}\hat{Y}_{\cdot j}^{a}$ $(i=1,2,\cdots,I)$ 로 주어진다. 여기에서 첨자 i는 대영역 내의 시군구 단위의 소지역들을 나타내고, 첨자 j는 성별(남, 여)-연령대별(15-34세, 35세 이상) 범주를 나타낸다.

η_{ij}는 i번째 소지역에서 j번째 성별-연령대별 범주에 대한 경제활동인구를 나타내며, $\eta_{ij}=(\zeta_{i0}^{C}/\zeta_{i0}^{R})\zeta_{ij}{}_{s}x_{j}$로 추정된다. 여기에서 ζ_{i0}^{C}는 2000년 상주추정인구, ζ_{i0}^{R}는 2000년 주민등록인구, ζ_{ij}는 2001년 5월(또는 6월)의 주민등록인구, ${}_{s}x_{j}$는 2001년 5월(또는 6월)의 경제활동인구조사자료에서 추정된 각

성별에 대한 경제활동 참가율을 나타낸다.

$\Psi^{a}_{\cdot j\cdot}$ 는 2001년 5월(또는 6월)의 경제활동인구조사에서 j번째 성별-연령대별 범주에 대한 실업률을 나타내며, $\Psi^{a}_{\cdot j\cdot} = \Upsilon_{\cdot j\cdot} / \sum_{i=1}^{I} \phi_{ij}$ $(j=1,2,3,4)$로 계산된다.

η_{ij}는 주민등록인구에 대한 상주 추정인구의 비로 표현된다. 필요에 따라 센서스 해인 1995년과 비교하여 추계 기준연도에 대한 상주 추정인구에 대한 벤치마킹이 요구된다. 그러나 2000년 광주 광역시 및 충청북도 내의 시군구 지역의 성별(남, 여)-연령대별(15-34세, 35세 이상) 주민등록인구에 대한 상주 추정인구의 비는 센서스 해인 1995년과 매우 유사한 경향을 보이므로 상주 추정인구에 대한 별도의 벤치마킹은 실시하지는 않았다.

i번째 소지역에서 j번째 성별-연령대별 범주에 대한 경제활동인구 η_{ij}를 상수로 가정한다면 합성 추정량 $\Upsilon_{i\cdot}^{S}$의 분산의 추정값은 식(5.6)을 통해 계산될 수 있다.

직접 추정량 $\Upsilon_{i\cdot}$은 합성 추정량 $\Upsilon_{i\cdot}^{S}$에 비해 표본크기에 민감하기 때문에 소지역 추정에서는 매우 큰 변동을 나타낸다. 또한 합성 추정량은 대영역에 대한 특성과 소지역에 대한 특성이 서로 상이할 경우에는 매우 큰 편향을 보일 것이다. 따라서 직접 추정량의 불안정성과 합성 추정량의 잠재적 편향 가능성을 보완하기 위해 두 추정량의 가중평균을 취한 복합 추정량 $\Upsilon_{i\cdot}^{C}$을 이용하여 실업자 총계를 추정하였다. 직접 추정량 $\Upsilon_{i\cdot}$과 합성 추정량 $\Upsilon_{i\cdot}^{S}$의 공분산이 0라는 가정 하에 복합 추정량 $\Upsilon_{i\cdot}^{C}$의 분산 추정은 식 (5.12)를 이용하여 계산할 수 있다.

5.3.3 Multi-level 모형을 이용한 계층적 베이즈 추정량

i번째 소지역에서 k번째 달에 실시한 경제활동인구조사에서 추정된 직접추정값을 y_{ik}라 하자. 이러한 직접추정값들이 Multi-level 모형을 이용한 계층적 베이즈 추정에서 종속변량으로 이용되었다. Multi-level 모형을 적합시켜 i번째 소지역에 대한 k번째 달의 실업자 수를 추정하기 위해서는 모형 적합에 필요한 적절한 설명변수와 반복적인 월별 자료들이 요구된다.

2001년 5월의 i번째 소지역에 대한 실업자 수를 추정하기 위해 2001년 2월부터 5월까지의 4개월 간의 경제활동인구조사 자료를 이용하였다. 2001년 6월의 i번째 소지역의 실업자 수 추정은 2001년 3월부터 6월까지의 4개월 간의 경제활동인구조사 자료를 이용하였다.

대영역인 광주광역시는 구단위의 소지역들로 이루어져 있다. 또한 대영역인 충청북도는 시단위의 소지역들과 군단위의 소지역들로 이루어져 있다. "Borrow Strength"를 적용하기 위해 5월(또는 6월) 경제활동인구조사 자료에서 광주광역시를 성별(남, 여)-연령대별(15-34세, 35세 이상)의 4개의 범주로 구분하여 각 범주의 실업률과 경제활동참가율을 계산한다.

한편, 충청북도는 시단위의 소지역들과 군단위의 소지역들로 구성되어 있으므로 시단위의 소지역들을 하나의 그룹으로 묶어 이에 대해 성별(남, 여)-연령대별(15-34세, 35세 이상)의 4개의 범주에 대한 실업률과 경제활동참가율을 계산하고, 군단위의 소지역들도 하나의 그룹으로 묶어 이에 대한 성별-연령대별의 4개 범주의 실업률과 경제활동참가율을 계산한다.

광주광역시 내의 i번째 소지역의 각 범주에 대한 해당 월의 경제활동인구는 각 범주에 대한 해당 월의 상주추정인구와 각 범주에 대한 해당 월의 경제활동참가율을 곱하여 얻는다.

i번째 소지역의 각 범주에 대한 해당 월의 실업자 수는 각 범주에 대한 해당 월의 경제활동인구에 각 범주의 실업률을 곱하여 얻는다. 충청북도에서도 같은 방법으로 i번째 소지역의 각 범주에 대한 경제활동인구와 실업자 수를 계산한다.

광주광역시 내의 i번째 소지역에서 k번째 월의 실업자 총계에 대한 직접추정값 y_{ik}는 위에서 계산된 i번째 소지역의 각 범주별 실업자 수를 설명변수로 하여 Multi-level 모형을 이용한 계층적 베이즈 구조하에서 보정된다.

선택된 설명변수 x_{ik}는 다음과 같이 주어진다.

$$x_{ik}=(x_{i1k}, x_{i2k}, x_{i3k}, x_{i4k})^T \ , \ k=1,2,3,4(\text{월})$$

$$=(\eta_{i1k}\Psi^{a}_{\cdot 1k\cdot}, \eta_{i2k}\Psi^{a}_{\cdot 2k\cdot}, \eta_{i3k}\Psi^{a}_{\cdot 3k\cdot}, \eta_{i4k}\Psi^{a}_{\cdot 4k\cdot})^T,$$

여기에서 첨자 i는 소지역, 첨자 j는 성별(남, 여)-연령대별(15-34세, 35세 이상) 범주, 첨자 k는 해당 월을 나타낸다. η_{ijk} ($j=1,2,3,4$)는 i번째 소지역에서 k번째 달에 실시한 경제활동인구조사에서 j번째 성별-연령대별 범주에 대한 경제활동인구를 나타내며, 이 값들은 2000년 주민등록인구와 상주추정인구에 의해 추정된다.

k번째 달에 실시한 경제활동인구조사에서 j번째 성별-연령대별 범주에 대한 실업률 $\Psi^{a}_{\cdot jk\cdot}$는 $\Psi^{a}_{\cdot jk\cdot} = Y_{\cdot jk\cdot} / \sum_{i=1}^{I} \phi_{ijk}$ ($j=1,2,3,4; k=1,2,3,4$)에 의해 계산된다. 여기에서 ϕ_{ijk}는 k번째 달에 실시한 경제활동인구조사에서 j번째 범주에 대한 경제활동인구를 나타낸다.

모형 (5.13)에서 i번째 소지역의 회귀계수 벡터 $\beta_i=(\beta_{i1},\beta_{i2},\beta_{i3},\beta_{i4})^T$는 다음과 같은 구조로 가정하였다.

$$y_{ik}=x_{i1k}\beta_{i1}+x_{i2k}\beta_{i2}+x_{i3k}\beta_{i3}+x_{i4k}\beta_{i4}+e_{ik}\ ,\ i=1,2,\cdots,I\ ;\ k=1,2,3,4$$

$$\beta_{i1}=\gamma_{10}+\nu_{i1}\ ;\ \beta_{i2}=\gamma_{20}+\nu_{i2}\ ;\ \beta_{i3}=\gamma_{30}+\nu_{i3}\ ;\ \beta_{i4}=\gamma_{40}+\nu_{i4}\ ,$$

여기에서 고정 회귀모수벡터 $\gamma=(\gamma_{10},\gamma_{20},\gamma_{30},\gamma_{40})^T$는 미지의 값이고 $\nu_i=(\nu_{i1},\nu_{i2},\nu_{i3},\nu_{i4})^T$는 결합분포 $\nu_i \overset{ind}{\sim} N_4(0\ ,\ \Phi)$를 가지며, 분산 공분산 행렬 Φ는 미지인 값이다. 오차항 e_{ik}는 평균 $E(e_{ik})=0$와 분산 $Var(e_{ik})=\sigma_i^2$를 갖는 서로 독립인 확률변수들이고, ν_i와 e_{ik}는 서로 독립이다.

i번째 소지역에서 k번째 달의 실업자 총계에 대한 계층적 베이즈 추정값과 추정값의 사후분산을 계산하기 위해 계층적 베이즈 Multi-level 모형 구조를 다음과 같이 설정하였다.

(i) $[y_{ik}|\ \beta_i\ ,\ \sigma_i^2] \overset{ind}{\sim} N(\ x_{ik}\ \beta_i\ ,\ \sigma_i^2)\ ,\ i=1,2,\cdots,I\ ;k=1,2,3,4$

(ii) $[\ \beta_i|\gamma,\ \Phi] \overset{ind}{\sim} N_4(Z_i\gamma,\ \Phi)\ ,\ i=1,2,\cdots,I$

(iii) 가정된 주변 확률사전분포는 다음과 같다:

$$\gamma\sim N_4(0\ ,\ D),\quad \tau_i \overset{ind}{\sim} G(a_i\ ,\ b_i),\quad \Omega\sim W_4(a\ ,\ R).$$

여기에서 $\tau_i=\sigma_i^{-2}$, $\Omega=\Phi^{-1}$을 나타내며, D, a_i, b_i, a, R은 주어지는 값이다. $G(a_i\ ,\ b_i)$는 밀도함수 $f(x)=\left[b_i^{a_i}/\Gamma(a_i)\right]x^{a_i-1}e^{-b_ix}$ $(a_i>0\ ,\ b_i>0\ ,\ x\geq 0)$를 갖는 감마분포이다.

i번째 소지역에서 k번째 달의 실업자 총계에 대한 사후추정치 $\mu_{ik} = \boldsymbol{x}_{ik}^T \boldsymbol{\beta}_i$ 는 주어진 $y=(\{y_{ik}\}, i=1,2,\cdots,I; k=1,2,3,4)$에 대한 $\boldsymbol{\beta}_i$ 의 사후분포로부터 생성된 충분한 양의 반복표본을 이용하여 계산될 수 있다.

이러한 표본 생성과정은 모수들에 대한 완전한 조건부 사후분포가 주어져야만 수행될 수 있다. 앞서 언급한 계층적 베이즈 Multi-level 모형 구조하에서 모수들에 대한 깁스 반복 표본을 생성하기 위해 필요한 조건부 사후분포들은 다음과 같다.

(i) $[\boldsymbol{\beta}_i | y, \gamma, \Omega, \tau] \overset{\text{ind}}{\sim} N_4\Big((\tau_i \sum_{k=1}^{4} \boldsymbol{x}_{ik} \boldsymbol{x}_{ik}^T + \Omega)^{-1} (\tau_i \sum_{k=1}^{4} y_{ik} \boldsymbol{x}_{ik} + \Omega Z_i \gamma), (\tau_i \sum_{k=1}^{4} \boldsymbol{x}_{ik} \boldsymbol{x}_{ik}^T + \Omega)^{-1} \Big)$,

(ii) $[\gamma | y, \beta, \Omega, \tau] \sim N_4\Big((\sum_{i=1}^{I} Z_i^T \Omega Z_i + D^{-1})(\sum_{i=1}^{I} Z_i^T \Omega \boldsymbol{\beta}_i), (\sum_{i=1}^{I} Z_i^T \Omega Z_i + D^{-1})^{-1} \Big)$,

(iii) $[\Omega | y, \beta, \gamma, \tau] \sim W_4\Big(\alpha + I, R + \frac{1}{2} \sum_{i=1}^{I} (\boldsymbol{\beta}_i - Z_i \gamma)(\boldsymbol{\beta}_i - Z_i \gamma)^T \Big)$,

(iv) $[\tau_i | y, \beta, \gamma, \Omega] \overset{\text{ind}}{\sim} G\Big(a_i + 2, b_i + \frac{1}{2} \sum_{k=1}^{4} (y_{ik} - \boldsymbol{x}_{ik}^T \boldsymbol{\beta}_i)^2 \Big)$.

위에 제시된 조건부 사후분포를 이용하여 모수들에 대한 반복표본을 생성한다. 알고리즘은 다음과 같다.

(step1) 초기값 $\gamma^{(0)}$, $\Omega^{(0)}$, $\tau_i^{(0)}$을 선택하여 (i)의 조건부 분포로부터 $\boldsymbol{\beta}_i$를 생성하고 이 생성표본을 $\boldsymbol{\beta}_i^{(1)}$이라 하자.

(step2) 초기값 $D^{(0)}$와 (step1)에서의 $\Omega^{(0)}$와 생성표본 $\boldsymbol{\beta}_i^{(1)}$을 이용하여

(ii)의 조건부 분포로부터 γ를 생성하고 이 생성표본을 $\gamma^{(1)}$이라 하자.

(step3) 초기값 $\alpha^{(0)}$, $R^{(0)}$와 (step2)에서의 $\beta_i^{(1)}$과 생성표본 $\gamma^{(1)}$을 이용하여 (iii)의 조건부 분포로부터 Ω를 생성하고 이 생성표본을 $\Omega^{(1)}$이라 하자.

(step4) 초기값 a_i, b_i와 (step3)에서의 $\beta_i^{(1)}$을 이용하여 (iv)의 조건부 분포로부터 τ_i를 생성하고 이 생성표본을 $\tau_i^{(1)}$이라 하자.

(step5) 위의 (step1)-(step4)의 과정을 반복하여 충분한 양의 모수들에 대한 깁스 반복표본을 생성한다.

위의 과정에서 생성된 모수들에 대한 반복표본 중 초기에 생성된 반복표본들은 불안정하므로 버리게 되는데 이 시점까지를 "Burn-in" 주기라 부른다. "Burn-in" 주기는 초기치 선택에 따라 길어질 수도 있고 짧아 질 수도 있다. "Burn-in" 주기 이후의 M개의 깁스 표본 $\{\beta_i^{(m)}, \gamma^{(m)}, \Omega^{(m)}, \tau_i^{(m)}; m=1,2,\cdots,M\}$을 $(\beta_i, \gamma, \Omega, \tau_i)$의 결합사후분포로부터 추출한 반복표본으로 간주한다. 여기에서 β_i의 사후추정값은 생성된 M개의 반복표본 $\{\beta_i^{(m)}; m=1,2,\cdots,M\}$들의 평균으로 계산된다.

깁스 반복 표본은 $D=diag(10^4, 10^4, 10^4, 10^4)$, $a=4$, $a=b=a_i=b_i=0.001$, R의 대각원소들은 1, 비대각원소들은 0.001로 하여 결정된 사전분포들을 이용하여 생하였다.

여기에서는 10,000개의 반복표본이 생성되었다. "Burn-in" 주기 이후의 5,000개(5,001~10,000)의 반복 표본을 이용하여 i번째 소지역의 실업자 총계에 대한 사후평균과 추정값의 분산을 계산하였다.

자료분석은 WinBUGS 프로그램(Spiegelhalter et al. 2000)을 이용하여 수행하였다.

5.3.4 시계열 및 횡단면 모형을 이용한 계층적 베이즈 추정량

i번째 소지역에서 t번째 달의 시계열 및 횡단면 모형을 이용한 계층적 베이즈 추정값은 2001년 2월부터 6월까지의 5개월 간의 경제활동인구조사 자료를 이용하여 추정하였다.

인접한 달의 실업자 추정값과의 추이를 살펴보기 위해 5월 실업자 총계 추정에서는 2001년 2월부터 5월까지의 연속적인 4개월 간의 경제활동인구조사 자료를 이용하였고, 6월 실업자 총계 추정에서는 2001년 3월부터 6월까지의 연속적인 4개월 간의 경제활동인구조사 자료를 이용하였다.

시계열 및 횡단면 자료 $\{y_{it}, \boldsymbol{x}_{it}\}$는 소지역 랜덤효과 ν_i와 $AR(1)$ 과정 u_{it}를 갖는 다음과 같은 선형혼합모형을 통해 보정된다.

$$y_{it} = \boldsymbol{x}_{it}^T \boldsymbol{\beta} + \nu_i + u_{it}\ ,\quad i=1,2,\cdots,I\ ;\ t=1,2,3,4$$

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it}\ ,\quad |\rho| < 1\ ,$$

여기에서 $\boldsymbol{\beta}=(\beta_1,\beta_2,\beta_3,\beta_4)^T$는 4×1 회귀계수 벡터, ν_i는 $\nu_i \overset{iid}{\sim} N(0, \sigma_i^2)$ 의 분포를 따르는 소지역 랜덤효과, u_{it}는 $AR(1)$ 과정, 오차항 $\varepsilon_{it} \overset{iid}{\sim} N(0, \sigma^2)$이며 $\{\nu_i\}$와 $\{\varepsilon_{it}\}$는 서로 독립임을 가정한다.

위의 모형에서 y_{it}는 광주광역시(또는 충청북도) 내의 i번째 소지역에서 t번째 달에 실시한 경제활동인구조사의 실업자 총계에 대한 직접추정값을 나타낸다.

Multi-level 모형을 이용한 계층적 베이즈 추정에서와 마찬가지로 월별 경제활동인구조사에서 대영역인 광주광역시(또는 충청북도)의 성별(남, 여)-연령대별(15-34세, 35세 이상) 4개 범주에 대한 경제활동참가율과 실업률을 "Borrow Strength"로 이용한다.

광주광역시(또는 충청북도) 내의 i번째 소지역의 t번째 달에 대한 각 범주별 경제활동인구는 t번째 달의 각 범주별 상주추정인구에 t번째 달의 해당 범주별 경제활동참가율을 곱하여 계산한다. i번째 소지역의 t번째 달에 대한 각 범주별 실업자 수는 t번째 달의 각 범주별 경제활동인구에 t번째 달의 해당 범주별 실업률을 곱하여 계산한다.

i번째 소지역에서 t번째 달의 실업자 총계에 대한 직접추정값 y_{it}를 보정하기 위해 선택된 횡단면 변량 $\boldsymbol{x}_{it}=(x_{i1t}, x_{i2t}, \cdots, x_{ipt})^T$는 다음과 같다.

$$\boldsymbol{x}_{it}=(x_{i1t}, x_{i2t}, x_{i3t}, x_{i4t})^T,\ \ i=1,2,\cdots,I\ ;\ \ t=1,2,3,4(\text{월})$$

$$=(\eta_{i1t}\Psi^{a}_{\cdot 1t\cdot}, \eta_{i2t}\Psi^{a}_{\cdot 2t\cdot}, \eta_{i3t}\Psi^{a}_{\cdot 3t\cdot}, \eta_{i4t}\Psi^{a}_{\cdot 4t\cdot})^T,$$

여기에서 첨자 i는 해당 소지역, 첨자 t는 해당 월을 나타낸다. η_{ijt} ($j=1,2,3,4$)는 t번째 달에 실시한 경제활동인구조사에서 i번째 소지역의 j번째 범주에 대한 경제활동인구를 나타낸다. t번째 달의 경제활동인구조사에서 j번째 성별-연령대별 범주의 실업률 $\Psi^{a}_{\cdot jt\cdot}$는 $\Psi^{a}_{\cdot jt\cdot}=\Upsilon_{\cdot jt\cdot}/\sum_{i=1}^{I}\phi_{ijt}$로 계산된다. $\Upsilon_{\cdot jt\cdot}$은 t번째 달에 실시한 경제활동인구조사에서 j번째 성별-연령대별 범주의 실업자 수를 나타내며, ϕ_{ijt}는 i번째 소지역에서 t번째 달

의 각 범주별 경제활동인구를 나타낸다.

위의 시계열 및 횡단면 모형을 이용한 계층적 베이즈 구조는 다음과 같이 가정하였다. 여기에서 $r_i=\sigma_i^{-2}$, $r=\sigma^{-2}$을 나타낸다.

(i) $[y_{it}|\nu_i,\beta,r_i,r,\rho] \overset{ind}{\sim} N\left(x_{it}^T\beta+\nu_i,\ \frac{r^{-1}}{1-\rho^2}\right)$, $k=1,2,3,4$

(ii) $[\nu_i|\beta,r_i,r,\rho] \overset{ind}{\sim} N(0,r_i^{-1})$, $i=1,2,\cdots,I$

(iii) 주변 사전확률분포는 다음과 같이 가정한다:

$\beta\sim Uniform(R^4)$, $r\sim G\left(\frac{a}{2},\frac{b}{2}\right)$, $r_i\sim G\left(\frac{c_i}{2},\frac{d_i}{2}\right)$, $\rho\sim f(\rho)$, 여기에서 ρ의 사전확률분포는 적절한 임의의 분포로 가정되며 $G(a,b)$는 확률밀도함수 $f(x)=[b^a/\Gamma(a)]x^{a-1}e^{-bx}$ $(a>0, b>0, x\geq 0)$를 갖는 감마분포이다.

광주광역시(또는 충청북도) 내의 i번째 소지역에서 t번째 달의 실업자 총계 $\mu_{it}=x_{it}^T\beta+\nu_i$의 사후추정값들을 계산하기 위해 깁스 반복 표본을 이용할 수 있다. 주어진 y에 대해서 μ_{it}의 사후평균 추정값들과 추정값들의 분산은 모수들에 대한 조건부 확률분포로부터 생성된 반복표본을 이용하여 계산한다.

깁스 반복표본을 생성하기 위해 필요한 모수들에 대한 조건부 확률분포는 다음과 같다.

(i) $[\beta|\nu_i,r_i,r,\rho,y]\sim N_4\Big((\sum_{i=1}^{I}\sum_{t=1}^{4}x_{it}x_{it}^T)^{-1}\sum_{i=1}^{I}\sum_{t=1}^{4}(y_{it}x_{it}-\nu_i x_{it}),$

$[r(1-\rho^2)]^{-1}(\sum_i\sum_t x_{it}x_{it}^T)^{-1}\Big)$,

(ii) $[\nu_i|\beta, r_i, r, \rho, y] \overset{ind}{\sim} N\Big([r(1-\rho^2)+r_i][r(1-\rho^2)]^{-1}\sum_{t=1}^{4}(y_{it}- x_{it}^T\beta),$

$[r(1-\rho^2)+r_i]^{-1}\Big)$

(iii) $[r_i|\beta, \nu_i, r, \rho, y] \overset{ind}{\sim} G\Big(\frac{1}{2}(c_i+\nu_i^2), \frac{1}{2}(d_i+1)\Big)$,

(iv) $[r|\beta, \nu_i, r_i, \rho] \sim G\Big(\frac{1}{2}\Big[a+(1-\rho^2)\sum_{i=1}^{I}\sum_{t=1}^{4}(y_{it}- x_{it}^T\beta-\nu_i)^2\Big]$,

$\frac{1}{2}(4I+b)\Big)$,

(v) $[\rho|\beta, \nu_i, r_i, r, y]=[A(\beta, \nu_i, r)]^{-1}(1-\rho^2)^{2I}f(\rho)$

$\times\exp\Big\{-\frac{r(1-\rho^2)}{2}\sum_{i=1}^{I}\sum_{t=1}^{4}(y_{it}- x_{it}^T\beta-\nu_i)^2\Big\}$,

단, $A(\beta, \nu_i, r)=\int(1-\rho^2)^{\frac{IT}{2}}\exp\Big\{-\frac{r(1-\rho^2)}{2}\sum_i\sum_t(y_{it}- x_{it}^T\beta-\nu_i)^2\Big\}$

$\times f(\rho)\,d\rho$.

초기값 $\beta^{(0)}$, $\nu_i^{(0)}$, $r_i^{(0)}$, $r^{(0)}$, $\rho^{(0)}$를 이용하여 위의 조건부 확률분포 (i)-(v)로부터 깁스 반복표본을 생성할 수 있다. 조건부 확률분포 (v)의 확률표본은 수치적분 $A(\beta, \nu_i, r)$을 직접적으로 계산하지 않고 채택 기각 표본추출 기법을 이용하여 반복 생성할 수 있다(Zeger and Karim, 1991). 특히 $\log A(\beta, \nu_i, r)$함수가 Concave 함수일 경우 베이지안 통계 분석의 전문프로그램인 WinBUGS를 이용하면 보다 안정적으로 (v)의 조건부분포에 대한 반복표본을 생성할 수 있다.

모수들에 대한 반복표본 생성 알고리즘은 다음 과정으로 수행된다.

(Step1) 초기값 $\nu_i^{(0)}$, $r^{(0)}$, $\rho^{(0)}$를 이용하여 (i)의 조건부 분포로부터 모수

β에 대한 표본을 생성하고 이 생성표본을 $\beta^{(1)}$라 하자.

(Step2) 초기값 $r_i^{(0)}$와 (Step1)에서의 $r^{(0)}$, $\rho^{(0)}$와 생성표본 $\beta^{(1)}$을 이용하여 (ii)의 조건부 분포로부터 모수 ν_i를 생성하고 이 생성표본을 $\nu_i^{(1)}$이라 하자.

(Step3) 초기값 $c_i^{(0)}$, $d_i^{(0)}$와 (Step2)에서의 생성표본 $\nu_i^{(1)}$을 이용하여 (iii)의 조건부 분포로부터 모수 r_i에 대한 표본을 생성하고 이 생성표본을 $r_i^{(1)}$이라 하자.

(Step4) 초기값 $a^{(0)}$, $b^{(0)}$와 (Step2)에서의 $\rho^{(0)}$, $\beta^{(1)}$와 (Step3)에서의 $\nu_i^{(1)}$을 이용하여 (iv)의 조건부 분포로부터 모수 r에 대한 표본을 생성하고 이 생성표본을 $r^{(1)}$이라 하자.

(Step5) 초기값 $\rho^{(0)}$와 (Step4)에서의 $\beta^{(1)}$, $\nu_i^{(1)}$, $r^{(1)}$을 이용하여 (v)의 조건부 분포로부터 모수 ρ에 대한 표본을 생성하고 이 생성표본을 $\rho^{(1)}$이라 하자.

(Step6) 위의 (Step1)-(Step5)의 과정을 반복하여 충분한 양의 반복표본을 생성한다.

위의 과정으로부터 "Burn-in" 주기 이후의 M개의 깁스 표본 $\{\beta^{(m)}, \nu_i^{(m)}, r_i^{(m)}, r^{(m)}, \rho^{(m)}; m=1,2,\cdots,M\}$을 모수 $(\beta, \nu_i, r_i, r, \rho)$의 결합확률분포로부터 생성된 반복표본으로 간주하여 i번째 소지역에서 t번째 달의 실업자 총계에 대한 추정이 이루어진다.

주어진 y에 대해 i번째 소지역의 t번째 달에 대한 실업자 총계

$\mu_{it} = x_{it}^T \beta + \nu_i$의 사후평균 추정값과 추정량의 분산은 M개의 깁스 반복 표본 $\{ \beta^{(m)}, \nu_i^{(m)}; m=1,2,\cdots,M\}$으로부터 일종의 MCMC 적분기법을 이용하여 계산한다.

깁스 반복 표본은 $a=b=c_i=d_i=0.002$, $\rho \sim Uniform(-1.0, 1.0)$로 하여 결정된 사전분포를 이용하여 생성하였다. 여기에서는 10,000개의 반복 표본이 생성되었다. "Burn-in" 주기 이후의 5,000개(5,001~10,000)의 반복 표본을 이용하여 2001년 5월과 6월의 i번째 소지역의 실업자 총계에 대한 사후평균과 추정값의 분산을 계산하였다. 자료분석은 WinBUGS 프로그램(Spiegelhalter et al. 2000)을 이용하여 수행하였다.

5.3.4 추정값들의 총계 보정

소지역 추정 방법에 의해 생산된 대영역 내의 소지역 추정값의 합은 반드시 대영역 총계 추정과 일치하지는 않는다. 대영역인 광주광역시 및 충북지역의 실업자 총계와 각 해당 지역 내의 소지역들에 대한 추정값의 합을 동일하게 나타내기 위해서 다음과 같은 비보정 추정절차를 적용하였다.

$$\hat{Y}_{i\cdot}^{A} = \frac{\hat{Y}_{i\cdot}}{\sum_{i=1}^{I} \hat{Y}_{i\cdot}} \hat{Y}^{*}, \quad i=1,2,\cdots,I,$$

여기에서 $\hat{Y}_{i\cdot}$은 합성 추정법, 복합 추정법, 계층적 베이즈 추정법에 의해 추정된 i번째 소지역의 실업자 수를 나타내며, $\hat{Y}^{*}$는 경제활동인구조사에서 직접 추정된 대영역에 대한 실업자 총계를 나타낸다.

5.3.5 추정 결과

(1) 대영역 표본설계 기반 추정(기존조사구 이용)

㈎ 광주광역시

2001년 5월과 6월의 경제활동인구조사를 기반으로 하여 추정된 광주 광역시 내의 시군구 단위의 소지역들에 대한 설계기반 및 모형기반 추정 결과가 다음 <표5.5>, <표5.6>, <표5.7>, <표5.8>에 주어졌다.

여기에서 RSE_i $(=CV_i)$는 i번째 시군구 단위의 소지역에 대한 상대표준오차를 나타내며 $(S.E/estimate)_i \times 100$ 의 값을 나타낸다.

다음 <표5.5>는 2001년 5월의 경제활동인구조사를 이용하여 설계기반 추정법인 직접 추정법, 합성 추정법과 복합 추정법을 적용하여 광주광역시의 각 구 단위 소지역에 대한 실업자 총계를 추정한 결과이다.

<표 5.5> 설계기반 소지역 추정값과 상대표준오차(2001년 5월)

시도	시군구	직접추정법			합성추정법			복합추정법		
		$\hat{Y}_{i\cdot}$	S.E	RSE	$\hat{Y}_{i\cdot}^{S}$	S.E	RSE	$\hat{Y}_{i\cdot}^{C}$	S.E	RSE
광주광역시	동구	2,807	976	**34.77**	2,285	281	10.65	2,397	270	10.18
	서구	4,837	1,164	24.06	5,168	627	10.50	5,165	552	9.66
	남구	5,289	1,137	21.50	4,253	518	10.54	4,499	471	9.46
	북구	8,629	1,256	14.56	8,602	1,046	10.53	8,498	804	8.55
	광산구	2,761	830	**30.06**	4,015	482	10.39	3,764	417	10.01
	합계	24,323			24,323			24,323		

광주광역시에서 구단위 소지역들의 직접 추정값의 상대표준오차(RSE)는 14.56%에서 34.77%의 범위에 있고, 합성 추정값의 상대표준오차는 10.39%에서 10.65%의 범위, 복합 추정값의 상대표준오차는 8.55%에서 10.18%의 범위에 있다.

캐나다의 소지역 통계는 상대표준오차의 목표요구정도를 약 25%정도로 인정하고 있다. 이를 기준으로 소지역 추정값의 신뢰성을 설명하기로 한다.

동구와 광산구의 직접추정값의 상대표준오차는 각각 34.77%와 30.06%로 신뢰성에 문제가 있다. 합성 추정값과 복합추정값의 상대표준오차는 모든 구 지역에서 25% 이내로 나타난다. 상대표준오차를 기준으로 볼 때 합성추정값들보다는 복합추정값들의 상대표준오차가 작게 나타난다.

따라서 광주광역시 내의 소지역 추정에서는 복합추정법이 직접추정법이나 합성추정법보다 효율적이다.

Multi-level 모형과 시계열 및 횡단면 모형을 이용한 계층적 베이즈 추정 결과는 다음 <표5.6>에 주어졌다.

<표5.6> 모형기반 소지역 추정값과 상대표준오차(2001년 5월)

시도	시군구	계층적 베이즈 추정법					
		Multi-level 모형			시계열/횡단면 자료 모형		
		$\hat{\mu}_{ik}^{HB_1}$	S.E	RSE	$\hat{\mu}_{it}^{HB_2}$	S.E	RSE
광주광역시	동구	3,166	437.2	14.70	2,802	479.1	17.21
	서구	4,969	830.3	17.79	5,391	512.3	9.56
	남구	4569	810.1	18.88	4,539	505.0	11.20
	북구	9,218	194.7	2.25	9,358	768.9	8.27
	광산구	2,399	454.6	20.18	2,233	723.0	**32.60**
	합계	24,323			24,323		

Multi-level 모형을 이용한 계층적 베이즈 추정값들의 상대표준오차는 2.25%~20.18%의 범위에 있고, 소지역 추정값들의 목표요구정도인 25%를 만족한다.

시계열 및 횡단면 모형을 이용한 계층적 베이즈 추정값들의 상대표준오차는 8.27%~32.60%의 범위에 있고, 특히 광산구의 상대표준오차는 32.60%로써 상대표준오차의 목표요구정도를 만족하지 못하며 다른 지역들은 모두 상대표준오차가 25%범위 내에 있다.

광주광역시의 5월 경제활동인구조사 결과에서는 설계기반 및 모형기반 소지역 추정값들 중 복합추정값들의 상대표준오차가 가장 작고 효율적이다.

다음 <표5.7>은 2001년 6월 설계기반 추정값들의 추정결과이다.

<표 5.7> 설계기반 소지역 추정값과 상대표준오차(2001년 6월)

시도	시군구	직접추정법			합성추정법			복합추정법		
		$\hat{Y}_{i\cdot}$	S.E	RSE	$\hat{Y}_{i\cdot}^{S}$	S.E	RSE	$\hat{Y}_{i\cdot}^{C}$	S.E	RSE
광주광역시	동구	2,331	602	**25.83**	2,685	364	12.67	2,854	311	11.40
	서구	1,819	1,251	**68.77**	6,209	855	12.87	5,345	706	13.82
	남구	13,336	3,324	24.93	5,054	694	12.84	6,004	680	11.85
	북구	8,986	1,496	16.65	10,257	1,407	12.82	10,508	1,025	10.21
	광산구	2,533	816	**32.21**	4,801	662	12.89	4,293	514	12.53
	합계	29,005			29,005			29,005		

직접추정값들의 상대표준오차는 16.65%~68.77%의 범위에 있고, 남구와 북구를 제외하고는 모두 소지역 추정값들의 상대표준오차가 목표요구정도인 25%를 벗어난다.

합성추정값들의 상대표준오차는 12.67%~12.89%의 범위에 있고 모든 지역들이 상대표준오차의 목표요구정도인 25%범위 내에 있다. 복합추정값들의 상대표준오차는 10.21%~13.82%의 범위에 있고, 모든 지역들에 대해 추정값들의 상대표준오차는 25%범위 내에 위치하며 상대표준오차의 효율 면에서는 합성추정값들보다 좋은 결과를 보여준다.

2001년 6월의 모형기반 추정값들의 추정결과는 다음 <표5.8>에 주어졌다.

<표5.8> 모형기반 소지역 추정값과 상대표준오차(2001년 6월)

시도	시군구	계층적 베이즈 추정법					
		Multi-level 모형			시계열/횡단면 자료 모형		
		$\hat{\mu}_{ik}^{HB_1}$	S.E	RSE	$\hat{\mu}_{ik}^{HB_2}$	S.E	RSE
광주광역시	동구	3,458	450.1	17.31	3,406	797.6	**26.26**
	서구	4,943	1311.0	**35.30**	5,848	1305.0	25.03
	남구	5,501	2878.0	**69.62**	6.074	1356.0	25.04
	북구	11,955	118.5	1.32	10,157	1946.0	21.49
	광산구	3,148	495.2	20.93	3,520	1232.0	**39.25**
	합계	29,005			29,005		

Multi-level 모형을 이용한 계층적 베이즈 추정값들의 상대표준오차는 1.32%~69.62%범위에 있고, 서구와 남구의 추정값들이 매우 불안정하며 소지역 추정값들의 상대표준오차의 목표요구정도인 25%를 만족하지 못한다.

시계열 및 횡단면 모형을 이용한 계층적 베이즈 추정값들의 상대표준오차는 21.49%~39.25%의 범위에 있고, 동구와 광산구의 추정값들의 상대표준오차가 목표요구정도를 만족하지 못하고 있다.

6월의 추정결과에서도 5월의 결과와 마찬가지로 복합추정법의 상대표준오차가 가장 안정적이며 다른 추정법들에 비해 효율이 좋다. 특히 모형기반 소지역 추정법은 일부 지역에 대해서는 다른 추정법보다 월등한 효율을 보이나 전체적으로 볼 때 상대표준오차의 변동이 심하며 상당히 불안정한 결과를 보인다.

(나) 충청북도

2001년 5월과 6월의 경제활동인구조사를 기반으로 하여 추정된 충청북도 내의 시군구 단위의 소지역들에 대한 추정 결과가 다음 <표5.9>, <표5.10>, <표5.11>, <표5.12>에 주어졌다.

다음 <표5.9>는 2001년 5월의 경제활동인구조사를 이용하여 설계기반 추정법인 직접 추정법, 합성 추정법과 복합 추정법을 적용하여 충청북도의 각 시군 단위 소지역에 대한 실업자 총계를 추정한 결과이다.

<표 5.9> 설계기반 소지역 추정값과 상대표준오차(2001년 5월)

시도	시군구	직접추정법			합성추정법			복합추정법		
		$\hat{Y}_{i\cdot}$	S.E	RSE	$\hat{Y}_{i\cdot}^{S}$	S.E	RSE	$\hat{Y}_{i\cdot}^{C}$	S.E	RSE
충청북도	청주시	9,929	1,948	19.62	9,341	1,248	11.51	9,434	1.051	9.94
	충주시	3,296	820	24.88	2,837	393	11.93	2,939	354	10.74
	제천시	751	475	**63.25**	1,798	250	11.97	1,604	221	12.29
	보은군	263	227	**86.31**	300	89	**29.28**	322	83	**27.76**
	옥천군	790	333	**42.15**	576	178	**30.53**	678	157	24.96
	영동군	252	189	**75.00**	577	178	**30.48**	461	130	**30.37**
	괴산군	779	649	**83.31**	295	88	**29.43**	331	87	**28.34**
	음성군	1,294	560	**43.28**	882	262	**29.34**	1,040	237	24.56
	청원군	767	307	**40.03**	1,343	413	**30.39**	1,054	246	25.15
	진천군	779	-	-	610	184	**29.82**	665	184	**29.82**
	단양군	0	-	-	341	102	**29.57**	372	102	**29.53**
	합계	18,900			18,900			18,900		

* 진천군은 표본 조사구 수가 1개이므로 직접추정값의 표준오차는 계산 불가함

5월 결과에서 직접추정값들의 상대표준오차는 19.62%~86.31%의 범위에 있고, 청주시와 충주시를 제외한 모든 시군지역들의 추정값에 대한 상대표준오차가 목표요구정도인 25%를 만족하지 못하고 있다.

따라서 충청북도지역 내의 시군단위 소지역에 대한 직접추정값들의 정도(Precision)는 신뢰할 수 없다.

합성추정값들의 상대표준오차는 11.51%~30.53%의 범위에 있고, 청주시, 충주시와 제천시의 시지역 추정값들을 제외하고 모든 군 지역들의 추정값들은 상대표준오차가 25%를 약간 벗어난 결과를 보인다.

복합추정값들의 상대표준오차는 9.94%~30.37%의 범위에 있고, 청주시, 충주시, 제천시를 포함하여 일부 군지역인 옥천군, 음성군과 청원군에 대한 추정값들의 상대표준오차가 25%이내에 있다. 기타 지역들은 추정값들의 상대표준오차가 25% 범위를 벗어나나 그 양은 미미하며, 합성추정값들보다는 상대표준오차의 양이 작고 보다 안정적이다.

2001년 5월의 충청북도 내의 시군지역들에 대한 모형기반 추정결과는 다음 <표5.10>에 주어졌다.

Multi-level 모형을 이용한 계층적 베이즈 추정값들의 상대표준오차는 8.04%~46.16%의 범위에 있고, 청주시, 제천시, 보은군, 괴산군과 음성군의 총 5개 시군지역들의 상대표준오차가 목표요구정도인 25% 이내에 있다.

시계열 및 횡단면 모형을 이용한 계층적 베이즈 추정값들의 상대표준오차는 11.38%~96.90%의 범위에 있고, 청주시, 충주시, 보은군, 옥천군, 괴산군, 음성군과 단양군의 총 7개 시군지역들의 상대표준오차가 25% 이내에 있

다. 한편, 제천시와 영동군의 추정값들의 상대표준오차는 각각 96.90%와 54.19%로 매우 불안정한 추정 결과를 보여준다.

<표5.10> 모형기반 소지역 추정값과 상대표준오차(2001년 5월)

시도	시군구	계층적 베이즈 추정법					
		Multi-level 모형			시계열/횡단면 자료 모형		
		$\hat{\mu}_{i*}^{HB_1}$	S.E	RSE	$\hat{\mu}_{it}^{HB_2}$	S.E	RSE
충청북도	청주시	10,196	810.2	8.04	10,150	1443.0	15.04
	충주시	3,093	964.0	**31.52**	2,937	636.0	22.90
	제천시	687	149.0	21.94	888	814.0	**96.90**
	보은군	298	69.6	23.59	303	78.9	22.73
	옥천군	1,052	303.9	**29.19**	1,187	154.8	11.38
	영동군	251	64.7	**26.08**	221	137.1	**54.19**
	괴산군	794	120.6	15.34	790	159.8	17.66
	음성군	1,109	178.7	16.28	986	180.6	15.98
	청원군	559	175.9	**31.81**	589	209.1	**30.98**
	진천군	402	183.7	**46.16**	483	147.4	**26.61**
	단양군	459	135.7	**29.89**	366	102.1	24.31
	합계	18,900			18,900		

충청북도내의 시군지역들에 대한 5월 추정결과에서는 복합추정법, Multi-level 모형을 이용한 계층적 베이즈 추정법과 시계열 및 횡단면 모형을 이용한 계층적 베이즈 추정법이 모두 유사한 추정 결과를 보이나, 추정값들의 안정성 측면에서는 복합추정법이 다른 추정법보다 좋은 결과를 나타낸다.

다음 <표5.11>은 2001년 6월의 설계기반 추정값들의 추정결과이다. 6월 결과에서 직접추정값들의 상대표준오차는 22.80%~117.9%의 범위에 있고, 청주시를 제외한 모든 충청북도 내의 시군지역들에 대한 추정값들의 상대표준오차는 목표요구정도인 25%를 만족하지 못하고 있다. 현행 표본조사구를 이용한 경제활동인구조사로는 소지역 추정에 직접추정법을 적용하는 것은 바람직하지 못하다.

<표 5.11> 설계기반 소지역 추정값과 상대표준오차(2001년 6월)

시도	시군구	직접추정법			합성추정법			복합추정법		
		$\hat{Y}_{i\cdot}$	S.E	RSE	$\hat{Y}_{i\cdot}^{S}$	S.E	RSE	$\hat{Y}_{i\cdot}^{C}$	S.E	RSE
충청북도	청주시	10,549	2,405	22.80	9,981	1,707	14.99	10,021	1,392	12.53
	충주시	3,297	858	**26.02**	3,041	524	15.10	3,088	447	13.05
	제천시	1,108	807	**72.83**	1,932	335	15.19	1,844	310	15.17
	보은군	0	0	-	226	61	25.31	235	61	25.31
	옥천군	263	207	**78.71**	436	120	**25.86**	404	104	25.12
	영동군	250	188	**75.20**	437	120	**25.81**	393	101	25.06
	괴산군	263	310	**117.9**	223	60	25.32	232	59	24.79
	음성군	1,828	668	**36.54**	667	180	25.35	767	174	22.14
	청원군	513	366	**71.35**	1,017	281	**25.95**	851	223	**25.57**
	진천군	513	0	-	462	126	**25.61**	480	126	**25.61**
	단양군	0	0	-	162	69	**39.88**	266	69	25.27
	합계	18,584			18,584			18,584		

합성추정값들의 상대표준오차는 14.99%~39.88%의 범위에 있고, 옥천군, 영동군, 청원군, 진천군과 단양군의 5개 군지역이 상대표준오차의 목표요구

정도를 만족하지 못한다.

복합추정값들의 상대표준오차는 12.53%~25.61%의 범위에 있고, 청원군과 진천군을 제외하고는 모든 시군지역들의 추정값들은 상대표준오차의 목표요구정도인 25% 이내에 있다. 또한 청원군과 진천군의 상대표준오차는 각각 25.57%와 25.61%로 목표요구정도에 근사하는 수치를 나타낸다.

2001년 6월의 모형기반 추정값들의 추정결과는 다음 <표5.12>에 주어졌다.

<표5.12> 모형기반 소지역 추정값과 상대표준오차(2001년 6월)

시도	시군구	계층적 베이즈 추정법					
		Multi-level 모형			시계열/횡단면 자료 모형		
		$\hat{\mu}_{ik}^{HB_1}$	S.E	RSE	$\hat{\mu}_{it}^{HB_2}$	S.E	RSE
충청북도	청주시	11,071	1447.0	12.74	10,799	658.8	5.97
	충주시	3,205	117.5	3.57	3,370	264.7	7.68
	제천시	678	293.6	**42.18**	786	515.6	**64.21**
	보은군	242	83.1	**38.12**	270	115.7	**31.96**
	옥천군	513	302.0	**65.37**	719	204.0	21.18
	영동군	259	31.0	13.32	185	213.2	**85.97**
	괴산군	427	171.6	**44.69**	389	198.6	**38.12**
	음성군	913	443.1	**53.91**	898	240.6	20.01
	청원군	579	181.7	**34.87**	362	244.0	**50.31**
	진천군	440	183.5	**46.34**	487	233.3	**35.78**
	단양군	257	132.9	**57.53**	319	140.0	**32.79**
	합계	18,584			18,584		

Multi-level 모형을 이용한 충북지역 시군단위 소지역 추정값들의 상대표준오차는 3.57%~65.37%의 범위에 있다. 청주시, 충주시와 영동군을 제외한 모든 시군단위 소지역들에 대한 추정값의 상대표준오차는 목표요구정도인 25%를 훨씬 벗어난 결과를 보인다.

시계열 및 횡단면 모형을 이용한 추정값들의 상대표준오차는 5.97%~85.97%의 범위에 있고, 청주시, 충주시, 옥천군과 음성군을 제외한 다른 시군단위 소지역들의 상대표준오차는 상대표준오차의 목표요구정도인 25%를 만족하지 못한다.

2001년 6월 추정결과에서는 다른 추정방법에 비해 복합추정법을 이용하여 추정한 결과가 시군단위 소지역들에 대한 상대표준오차가 가장 작고 보다 안정적이다.

(4) 소지역 추정법 대비 상대표준오차 분포

다음 <표5.13>은 이상에서 논의된 5가지 추정법들에 대해서 광주광역시 내의 구단위 소지역 추정값들의 상대표준오차에 대한 분포를 요약한 것이다. 여기에 집계된 수는 2001년 5월과 6월 결과에서 상대표준오차 범위에 있는 소지역들의 수이다.

소지역 추정값들의 상대표준오차의 목표요구정도를 25%로 정하고 광주광역시의 구단위 소지역들의 실업자 총계를 추정할 때, 직접추정법은 약 50%의 소지역들의 추정값이 상대표준오차의 목표요구정도를 만족한다.

합성추정법과 복합추정법은 전체 100%의 소지역들에 대한 추정값의 상대표준오차가 목표요구정도를 만족하나, 복합추정법이 합성추정법보다는 안

정적이며 효율적이다.

모형기반 추정법에서는 Multi-level 모형을 이용한 계층적 베이즈 추정법은 약 80%의 소지역들에 대한 추정값의 상대표준오차가 목표요구정도를 만족하며, 시계열 및 횡단면 모형을 이용한 계층적 베이즈 추정법은 약 70%의 소지역들이 상대표준오차의 목표요구정도를 만족한다.

<표5.13> 광주광역시에 대한 상대표준오차 분포 비교(5, 6월 결과)

구 분			상대표준오차(%)				
			5.4% 이하	5.4 ~15.4%	15.5 ~25.4%	25.5 ~35.4%	35.5% 이상
설계 기반 추정	직접추정	소지역 수	0	2	3	4	1
		누율(%)	0.0	20.0	50.0	90.0	100.0
	합성추정	소지역 수	0	10	0	0	0
		누율(%)	0.0	100.0	100.0		
	복합추정	소지역 수	0	10	0	0	0
		누율(%)	0.0	100.0	100.0		
모형 기반 추정	Multi-level 모형 베이즈 추정	소지역 수	2	1	5	1	1
		누율(%)	20.0	30.0	80.0	90.0	100.0
	시계열 및 횡단면 모형 베이즈추정	소지역 수	0	3	4	2	1
		누율(%)	0.0	30.0	70.0	90.0	100.0

각 추정방법에 대한 상대표준오차의 분포를 비교해 볼 때 위의 5가지 추정법 중 복합추정법을 이용한 시군구단위 소지역 추정결과 다른 추정법들에 비해 훨씬 안정적이며 효율적이다.

다음 <표5.14>는 충청북도 내의 시군단위 소지역 추정값들의 상대표준오차에 대한 분포를 요약한 것이다. 여기에 집계된 수는 2001년 5월과 6월의 두 달간의 추정결과에서 상대표준오차 범위에 있는 소지역들의 수를 집계한 것이다.

<표5.14> 충청북도에 대한 상대표준오차 분포 비교(5, 6월 결과)

구 분			상대표준오차(%)				
			5.4% 이하	5.4 ~15.4%	15.5 ~25.4%	25.5 ~35.4%	35.5% 이상
설계 기반 추정	직접추정	소지역 수	0	0	3	1	13
		누율(%)	0.0	0.0	17.6	23.5	100.0
	합성추정	소지역 수	0	6	3	12	1
		누율(%)	0.0	27.3	40.9	95.5	100.0
	복합추정	소지역 수	0	6	9	7	0
		누율(%)	0.0	27.3	68.2	100.0	
모형 기반 추정	Multi-level 모형 베이즈 추정	소지역 수	1	4	3	6	8
		누율(%)	4.5	22.7	36.4	63.6	100.0
	시계열 및 횡단면 모형 베이즈추정	소지역 수	0	4	7	4	7
		누율(%)	0.0	18.2	50.0	68.2	100.0

소지역 추정값들의 상대표준오차의 목표요구정도를 25%로 정하고 충청북도의 시군단위 소지역들의 실업자 총계를 추정할 때, 직접추정법은 약 17.6%의 소지역들의 추정값이 상대표준오차의 목표요구정도를 만족한다.

합성추정법은 약 40.9%의 소지역들에 대한 추정값의 상대표준오차가 상대표준오차의 목표요구정도를 만족하며, 복합추정법은 약 68.2%의 소지역들이 상대표준오차의 목표요구정도를 만족한다. 광주광역시에와 마찬가지로 복합추정법이 합성추정법보다는 안정적이며 효율적이다.

모형기반 추정법에서는 Multi-level 모형을 이용한 계층적 베이즈 추정법은 약 36.4%의 소지역들에 대한 추정값의 상대표준오차가 목표요구정도를 만족하며, 시계열 및 횡단면 모형을 이용한 계층적 베이즈 추정법은 약 50%의 소지역들이 상대표준오차의 목표요구정도를 만족한다.

각 추정방법에 대한 상대표준오차의 분포를 비교해 볼 때 위의 5가지 추정법 중 복합추정법을 이용한 소지역 추정결과가 다른 추정법들에 비해 훨씬 안정적이며 효율적이다.

(2) 소지역 표본조사구 수 증편에 따른 추정

광주광역시 및 충청북도 시군구 단위의 소지역들에 대한 직접 추정값의 신뢰성을 확보할 목적으로 각 시군구에 대한 표본 증편이 이루어졌다. 증편된 표본 조사구 수의 현황은 다음 <표5.15>에 주어졌다.

광주광역시는 기존의 87개 조사구에 14개의 표본조사구가 추가되어 총 101개의 조사구로 증편되었고, 충청북도는 기존의 64개 조사구에 74개의 조사구가 추가되어 총 138개의 조사구로 증편되었다. 증편된 표본 조사구를 이용하여 2001년 5월과 6월에 걸쳐 광주광역시 및 충청북도에 대한 경제활동인구조사가 실시되었다.

<표 5.15> 증편된 표본 조사구 수 현황

시도	시군구	표본 조사구 수		
		기존	증편	증가량
광주광역시	동구	10	14	+4
	서구	15	18	+3
	남구	16	18	+2
	북구	34	34	0
	광산구	12	17	+5
	합계	87	101	+14
충청북도	청주시	22	22	0
	충주시	11	16	+5
	제천시	4	14	+10
	보은군	2	10	+8
	옥천군	3	10	+7
	영동군	3	10	+7
	괴산군	5	11	+6
	음성군	6	11	+5
	청원군	5	14	+9
	진천군	1	10	+9
	단양군	2	10	+8
	합계	64	138	+74

(나) 광주광역시

시군구 단위 소지역들에 대한 설계기반 추정량들의 추정 결과는 다음 <표5.16>과 <표5.17>에 주어졌다. 모형기반 추정량들은 최소한 3개월 이상

의 층편조사구를 이용한 조사결과가 있어야 추정이 가능하므로 여기에서는 언급하지 않는다.

여기에서 RSE_i ($=CV_i$)는 i번째 시군구 단위의 소지역에 대한 상대표준오차를 나타내며 $(S.E/estimate)_i \times 100$의 값을 나타낸다.

다음 <표5.16>은 2001년 5월에 층편된 표본조사구를 적용하여 각 시군구 단위의 소지역에 대한 실업자 총계를 추정한 결과이다.

<표5.16> 설계기반 추정값과 상대표준오차(2001년 5월)

시도	시군구	직접추정법			합성추정법			복합추정법		
		$\hat{Y}_{i\cdot}$	S.E	RSE	$\hat{Y}_{i\cdot}^{S}$	S.E	RSE	$\hat{Y}_{i\cdot}^{C}$	S.E	RSE
광주광역시	동구	3,403	968	27.37	2,293	276	9.98	2,505	266	9.42
	서구	4,922	1,188	23.23	5,160	617	9.91	5,312	548	9.15
	남구	5,818	1,100	18.19	4,259	509	9.91	4,699	462	8.72
	북구	7,180	1,086	14.55	8,588	1,029	9.93	7,973	747	8.31
	광산구	3,000	717	23.00	4,023	474	9.77	3,834	396	9.16
	합계	24,323			24,323			24,323		

광주광역시에서 구단위 소지역들의 직접 추정값의 상대표준오차는 14.55%~27.37%의 범위에 있고, 동구를 제외하고 모든 소지역들에 대한 추정값의 상대표준오차는 25% 이내에 있다. 합성 추정값의 상대표준오차는

9.91%~9.98%의 범위에 있고, 복합 추정값의 상대표준오차는 8.32%~9.42%의 범위에 있다.

합성추정값과 복합추정값들에 대한 상대표준오차는 모두 상대표준오차의 목표요구정도인 25% 이내를 만족하며, 복합추정값들의 상대표준오차가 합성추정값들보다 약간 작은 값을 나타낸다.

세가지 추정법 중 복합추정법의 효율이 가장 좋다.

다음 <표5.17>은 층편된 표본조사구를 적용하여 2001년 6월의 광주광역시 및 충청북도 내의 각 시군구 단위 소지역에 대한 실업자 총계를 추정한 결과이다.

<표5.17> 설계기반 추정값과 상대표준오차(2001년 6월)

시도	시군구	직접추정법			합성추정법			복합추정법		
		$\hat{Y}_{i\cdot}$	S.E	RSE	$\hat{Y}_{i\cdot}^{S}$	S.E	RSE	$\hat{Y}_{i\cdot}^{C}$	S.E	RSE
광주광역시	동구	3,465	542	18.36	2,705	262	9.59	2,908	236	8.51
	서구	6,612	1,029	18.27	6,197	609	9.73	6,396	524	8.59
	남구	6,597	1,219	21.69	5,052	492	9.64	5,430	456	8.81
	북구	9,128	1,295	16.65	10,252	1,009	9.74	9,842	796	8.48
	광산구	3,203	722	**26.46**	4,799	467	9.63	4,429	392	9.28
	합계	29,005			29,005			29,005		

6월 층편조사구를 적용한 구단위 소지역들에 대한 직접 추정값의 상대표준오차는 16.65%~26.46%의 범위에 있고, 광산구를 제외한 모든 소지역들에

대한 추정값들의 상대표준오차는 상대표준오차의 목표요구정도인 25%이내에 있다.

합성 추정값의 상대표준오차는 9.59%~9.74%의 범위에 있고, 복합 추정값의 상대표준오차는 8.48%~9.28%의 범위에 있다. 모든 지역들에 대해서 합성추정값과 복합추정값들의 상대표준오차가 목표요구정도를 만족하며, 특히 복합 추정법의 효율이 다른 추정법들에 비해 좋게 나타난다.

(나) 충청북도

층편된 표본 조사구를 이용하여 2001년 5월과 6월에 걸쳐 광주광역시 및 충청북도에 대한 경제활동인구조사가 실시되었다. 시군구 단위 소지역들에 대한 설계기반 추정량들의 추정 결과는 다음 <표5.18>과 <표5.19>에 주어졌다.

다음 <표5.18>은 2001년 5월에 층편된 표본조사구를 적용하여 충청북도 내의 각 시군 단위의 소지역에 대한 실업자 총계를 추정한 결과이다.

<표5.18> 설계기반 추정값과 상대표준오차(2001년 5월)

시도	시군구	직접추정법			합성추정법			복합추정법		
		$\hat{Y}_{i\cdot}$	S.E	RSE	$\hat{Y}_{i\cdot}^{S}$	S.E	RSE	$\hat{Y}_{i\cdot}^{C}$	S.E	RSE
충청북도	청주시	8,357	1,449	19.58	9,351	1,083	10.42	9,036	867	9.30
	충주시	3,084	731	**26.77**	2,831	338	10.74	2,980	307	9.99
	제천시	2,535	598	**26.64**	1,794	213	10.68	1,960	201	9.94
	보은군	749	183	23.05	307	44	15.55	341	43	13.83
	옥천군	837	224	25.25	568	85	16.19	625	80	14.04
	영동군	250	148	**55.85**	567	86	16.41	504	74	16.12
	괴산군	503	349	**65.48**	301	43	15.47	309	43	15.30
	음성군	661	329	**46.93**	898	128	15.42	893	119	14.64
	청원군	918	319	**32.79**	1,324	198	16.19	1,267	168	14.56
	진천군	584	300	**48.47**	613	89	15.70	627	85	14.89
	단양군	422	208	**46.53**	346	50	15.63	358	48	14.72
	합계	18,900			18,900			18,900		

시군단위의 소지역들에 대한 직접 추정값의 상대표준오차는 19.58%~65.48%의 범위에 있고, 청주시, 보은군과 옥천군을 제외한 나머지 시군단위 소지역들에 대한 추정값의 상대표준오차는 상대표준오차의 목표요구정도인 25%를 만족하지 못한다.

합성 추정값들의 상대표준오차는 10.42%~16.41%의 범위에 있고, 복합 추정값의 상대표준오차는 9.30%~16.12%의 범위에 있다. 합성추정법과 복합 추정법으로 추정한 추정값들은 모든 지역들에 대해서 상대표준오차의 목표 요구정도를 만족한다.

세가지 방법에 의한 추정값들 중 복합 추정값들의 상대표준오차가 가장 작다.

5월 경제활동인구조사결과와 마찬가지로 추가조사구를 적용한 경제활동 인구조사 결과에서도 복합 추정법의 효율이 가장 좋게 나타난다.

다음 <표5.19>는 증편된 표본조사구를 적용하여 2001년 6월의 충청북도 내의 각 시군 단위 소지역에 대한 실업자 총계를 추정한 결과이다.

<표5.19> 설계기반 추정값과 상대표준오차(2001년 6월)

시도	시군구	직접추정법			합성추정법			복합추정법		
		$\hat{Y}_{i\cdot}$	S.E	RSE	$\hat{Y}_{i\cdot}^{S}$	S.E	RSE	$\hat{Y}_{i\cdot}^{C}$	S.E	RSE
충청북도	청주시	8,985	1,806	22.81	9,984	1,475	13.51	9,659	1,142	11.75
	충주시	3,112	748	**27.28**	3,041	450	13.54	3,151	386	12.18
	제천시	2,857	794	**31.55**	1,929	286	13.56	2,144	269	12.48
	보은군	425	271	**51.42**	227	35	16.36	231	35	15.91
	옥천군	498	198	**32.09**	414	68	17.48	433	64	15.50
	영동군	281	236	**67.82**	411	68	17.57	403	66	17.19
	괴산군	501	400	**64.41**	224	35	16.67	224	35	16.43
	음성군	641	216	**27.17**	670	104	16.51	694	93	14.07
	청원군	714	332	**37.51**	970	156	17.12	951	141	15.56
	진천군	498	240	**38.90**	457	71	16.55	466	68	15.32
	단양군	72	92	**103.4**	257	40	16.60	228	37	17.05
	합계	18,584			18,584			18,584		

충청북도 시군 단위의 소지역들에 대한 직접 추정값들의 상대표준오차는 22.81%~103.40%의 범위에 있고, 청주시를 제외한 모든 소지역 추정값들에 대한 추정값의 상대표준오차는 목표요구정도인 25%를 훨씬 벗어나 있다.

합성 추정값들의 상대표준오차는 13.51%~17.57%의 범위에 있고, 복합 추정값의 상대표준오차는 11.75%~17.19%의 범위에 있다.

합성추정값들과 복합추정값들은 모든 소지역들에 대해서 상대표준오차의

목표요구정도를 만족한다. 세가지 추정값들 중 복합 추정값의 상대표준오차가 가장 작다.

6월 경제활동인구조사 결과와 마찬가지로 추가조사구를 적용한 경제활동인구조사 결과에서도 복합 추정법의 효율이 가장 좋게 나타난다.

㈐ 소지역 추정법 대비 상대표준오차 분포

다음 <표5.20>은 직접추정법, 합성추정법과 복합추정법에 대해서 증편조사구 적용 시 광주광역시 내의 구단위 소지역 추정값들의 상대표준오차에 대한 분포를 요약한 것이다.

여기에 집계된 수는 2001년 5월과 6월 증편조사구를 적용한 경제활동인구조사 추정결과에서 상대표준오차 범위에 있는 소지역들의 수이다.

소지역 추정값들의 상대표준오차의 목표요구정도를 25%로 정하고 광주광역시의 구단위 소지역들의 실업자 총계를 추정할 때, 직접추정법은 약 80%의 소지역들이 추정값의 상대표준오차에 대한 목표요구정도를 만족한다.

기존 조사구를 이용했을 경우 약 50%의 소지역들이 추정값의 상대표준오차의 목표요구정도를 만족한데 비하면 증편조사구 적용 시 약 30%의 소지역들이 추가적으로 상대표준오차의 목표요구정도를 만족하는 것으로 나타났다.

<표5.20> 광주광역시 상대표준오차 분포 비교(5, 6월 증편조사 결과)

구 분			상대표준오차(%)				
			5.4% 이하	5.4 ~15.4%	15.5 ~25.4%	25.5 ~35.4%	35.5% 이상
기존 조사	직접추정	소지역 수	0	2	3	4	1
		누율(%)	0.0	20.0	50.0	90.0	100.0
	합성추정	소지역 수	0	10	0	0	0
		누율(%)	0.0	100.0	100.0		
	복합추정	소지역 수	0	10	0	0	0
		누율(%)	0.0	100.0	100.0		
증 편 조사	직접추정	소지역 수	0	1	7	2	0
		누율(%)	0.0	10.0	80.0	100.0	
	합성추정	소지역 수	0	10	0	0	0
		누율(%)	0.0	100.0	100.0		
	복합추정	소지역 수	0	10	0	0	0
		누율(%)	0.0	100.0	100.0		

합성추정법과 복합추정법은 증편조사구 적용 시 기존조사 결과와 마찬가지로 전체 100%의 소지역들에 대한 추정값의 상대표준오차가 목표요구정도를 만족한다.

다음 <표5.21>은 증편조사구 적용 시 충청북도 내의 시군단위 소지역 추정값들의 상대표준오차에 대한 분포를 요약한 것이다. 여기에 집계된 수는 2001년 5월과 6월의 두 달간의 추정결과에서 상대표준오차 범위에 있는 소지역들의 수를 집계한 것이다.

<표5.21> 충청북도 상대표준오차 분포 비교(5, 6월 증편조사 결과)

구 분			상대표준오차(%)				
			5.4% 이하	5.4 ~15.4%	15.5 ~25.4%	25.5 ~35.4%	35.5% 이상
기존 조사	직접추정	소지역 수	0	0	3	1	13
		누율(%)	0.0	0.0	17.6	23.5	100.0
	합성추정	소지역 수	0	6	3	12	1
		누율(%)	0.0	27.3	40.9	95.5	100.0
	복합추정	소지역 수	0	6	9	7	0
		누율(%)	0.0	27.3	68.2	100.0	
증편 조사	직접추정	소지역 수	0	0	4	7	11
		누율(%)	0.0	0.0	18.1	50.0	100.0
	합성추정	소지역 수	0	8	14	0	0
		누율(%)	0.0	36.3	100.0		
	복합추정	소지역 수	0	15	7	0	0
		누율(%)	0.0	68.2	100.0		

증편조사구 적용 시 직접추정값들의 상대표준오차의 목표요구정도인 25% 이내에 해당하는 소지역들은 전체 소지역의 약 18.1%로 기존조사 결과와 별 차이가 없다.

반면 합성추정법 및 복합추정법을 이용하여 추정하였을 경우 충청북도 내의 모든 소지역들에 대한 추정값의 상대표준오차가 상대표준오차의 목표요구정도인 25% 이내를 만족한다. 증편조사구 적용 시 합성추정법은 전체 소지역들 중 약 59.1%, 복합추정법은 약 31.8%가 추가로 상대표준오차의 목표요구정도를 만족하는 것으로 나타났다.

(3) 소지역 표본조사구 수 증편에 따른 효율비교

시군구 단위의 소지역들에 대한 표본 조사구 수 증편에 따른 소지역 추정값들의 효율을 증편 전과 비교하였다. 기존 경제활동인구조사에서 조사된 소지역 추정값들의 상대표준오차(*RSE*)와 표본 조사구 수가 증편된 상태에서 실사된 소지역 추정값들의 상대표준오차의 차이를 이용하여 상대효율의 이득을 계산하였다.

(가) 광주광역시

2001년 5월과 6월 경제활동인구조사에 대한 설계기반 추정량들의 상대효율 이득에 관련한 결과가 다음 <표5.22>와 <표5.23>에 주어졌다. 여기에서 상대효율 이득은 표본 조사구 증편에 따른 상대표준오차의 절감량을 백분율로 나타낸 값이며 다음 식과 같이 주어진다.

$$\text{상대효율 이득} = \frac{RSE(\text{기존}) - RSE(\text{증편})}{RSE(\text{기존})} \times 100\ (\%)$$

<표5.22>에서 5월의 경제활동인구조사에서 각각의 추정법들에 대한 상대효율 이득을 표본조사구 수 증편 전과 증편 후로 나누어 비교하였다.

광주광역시의 구단위 소지역들에 대한 표본조사구 수 추가에 따른 기존 직접추정값 대비 증편 조사의 직접추정값들의 효율이득은 최소 0.07%에서 최대 23.49%의 범위에 있고 평균 12.74%의 효율이득이 발생하였다. 기존 합성추정값 대비 증편조사 합성추정추정값들의 상대효율이득은 평균 5.91%의 효율이득이 발생하였다. 기존 복합추정값 대비 증편조사 복합추정값들의 상대효율이득은 평균 6.37%의 효율이득이 발생하였다.

<표5.22> 설계기반 추정량들의 상대효율 이득(2001년 5월)

시도	시군구	직접추정량			합성추정량			복합추정량		
		RSE (기존 조사)	RSE (증편 조사)	효율 이득 (%)	RSE (기존 조사)	RSE (증편 조사)	효율 이득 (%)	RSE (기존 조사)	RSE (증편 조사)	효율 이득 (%)
광주광역시	동구	34.77	27.37	21.28	10.65	9.98	6.29	10.18	9.42	7.47
	서구	24.06	23.23	3.45	10.50	9.91	5.62	9.66	9.15	5.28
	남구	21.50	18.19	15.40	10.54	9.91	5.98	9.46	8.72	7.82
	북구	14.56	14.55	0.07	10.53	9.93	5.70	8.55	8.31	2.81
	광산구	30.06	23.00	23.49	10.39	9.77	5.97	10.01	9.16	8.49
	평균			12.74			5.91			6.37

6월의 경제활동인구조사에서 표본 조사구 수 추가에 따른 각 추정량들의 효율이득은 다음 <표5.23>에 주어졌다.

<표5.23> 설계기반 추정량들의 상대효율 이득(2001년 6월)

시도	시군구	직접추정량			합성추정량			복합추정량		
		RSE (기존 조사)	RSE (증 편 조사)	효율 이득 (%)	RSE (기존 조사)	RSE (증편 조사)	효율 이득 (%)	RSE (기존 조사)	RSE (증편 조사)	효율 이득 (%)
광주광역시	동구	25.83	18.36	28.92	12.67	9.59	24.31	11.40	8.51	25.35
	서구	**68.77**	18.27	73.43	12.87	9.73	24.40	13.82	8.59	37.84
	남구	24.93	21.69	13.00	12.84	9.64	24.92	11.85	8.81	25.65
	북구	16.65	16.65	0.00	12.82	9.74	24.02	10.21	8.48	16.94
	광산구	**32.21**	26.46	17.85	12.89	9.63	25.29	12.53	9.28	25.94
	평균			26.64			24.59			26.34

6월 구단위 소지역들에 대한 기존 직접 추정값 대비 증편조사 직접추정값들의 효율이득은 최소 0.00%에서 최대 73.43%의 범위에 있고 평균 26.64%의 효율이득이 발생하였다.

기존 합성 추정값 대비 증편조사 합성추정값들의 상대효율이득은 평균 24.59%의 효율이득이 발생하였다. 기존 복합추정값 대비 증편조사 복합 추정값들의 상대효율이득은 평균 26.34%의 효율이득이 발생하였다.

다음 <표5.24>와 <표5.25>에서는 기존 직접추정값 대비 증편조사 직접 추정값들의 상대효율이득과 기존 직접추정값 대비 기존 설계기반 추정값 및

기존 모형기반 추정값들의 상대효율이득을 요약한 결과이다.

상대효율 이득은 다음 식으로 계산된다.

$$\text{상대효율 이득} = \frac{\text{직접추정값의 } RSE - \text{기타 추정값의 } RSE}{\text{직접추정값의 } RSE} \times 100\,(\%)$$

다음 <표5.24>는 2001년 5월의 경제활동인구조사 자료를 이용한 결과이다.

기존 직접추정값 대비 중편조사 직접추정값들의 상대효율이득은 최소 0.07%에서 최대 23.49%의 범위에 있고, 조사구 수 증편 전에 비해 평균 12.74%의 상대효율이득이 발생하였다.

기존 직접추정값 대비 기존 합성 추정값들의 상대효율 이득은 평균 53.96%, 기존 직접추정값 대비 기존 복합 추정값들의 상대효율이득은 평균 58.91%의 효율이득이 발생하였다. 기존 직접추정값 대비 기존 Multi-level 모형을 이용한 계층적 베이즈 추정값들의 상대효율이득은 평균 42.68%, 시계열 및 횡단면 모형을 이용한 계층적 베이즈 추정값들의 상대효율이득은 평균 38.68%의 효율이득이 발생하였다.

<표5.24> 기존 직접추정값 대비 상대효율 이득 비교(2001년 5월)

시도	시군구	직접추정 (기존조사구 이용)		직접추정 (증편조사구 이용)			합성추정 (기존조사구 이용)			복합추정 (기존조사구 이용)		
		추정값	RSE	추정값	RSE	효율 이득 (%)	추정값	RSE	효율 이득 (%)	추정값	RSE	효율 이득 (%)
광주광역시	동구	2,807	**34.77**	3,403	27.37	21.28	2,285	10.65	69.37	2,397	10.18	70.72
	서구	4,837	24.06	4,922	23.23	3.45	5,168	10.50	56.36	5,165	9.66	59.85
	남구	5,289	21.50	5,818	18.19	15.40	4,253	10.54	50.98	4,499	9.46	56.00
	북구	8,629	14.56	7,180	14.55	0.07	8,602	10.53	27.68	8,498	8.55	41.28
	광산구	2,761	30.06	3,000	23.00	23.49	4,015	10.39	65.44	3,764	10.01	66.70
	합계	24,323		24,323		(12.74)	24,323		(53.96)	24,323		(58.91)

시도	시군구	Multi-level 모형 추정량 (기존조사구 이용)			시계열 및 횡단면 모형 추정량 (기존조사구 이용)		
		추정값	RSE	효율 이득 (%)	추정값	RSE	효율 이득 (%)
광주광역시	동구	3,168	437.2	57.73	2,802	479.1	50.51
	서구	4,969	830.2	26.06	5,391	512.3	60.25
	남구	4,569	810.1	12.18	4,539	505.0	47.91
	북구	9,218	194.7	84.55	9,358	768.9	43.18
	광산구	2,399	454.6	32.88	2,233	723.0	-8.43
	합계	24,323		(42.68)	24,323		(38.68)

복합추정값들에 대한 상대효율이득이 타 추정방법에 의한 추정값들의 상

대효율이득보다 월등히 큰 값을 나타낸다. 표본 조사구 증편 후의 직접추정값들에 대한 상대효율이득은 평균 12.74%로써 기존조사를 이용한 복합추정값들의 평균 상대효율이득인 58.91%에 크게 못 미친다. 따라서 복합추정법은 표본조사구 수가 적을 경우에도 안정성 및 효율측면에서 다른 추정법에 비해 월등히 좋은 추정결과를 보인다.

다음 <표5.25>는 2001년 6월의 경제활동인구조사 자료를 이용한 광주광역시 구단위 소지역들의 추정결과이다.

6월 경제활동인구조사 결과에서 기존 직접추정값 대비 증편조사 직접추정값들의 상대효율이득은 최소 0.00%에서 최대 73.43%의 범위에 있고, 증편조사로부터 평균 26.64%의 효율이득이 발생하였다.

기존 직접추정값 대비 기존 합성추정값들의 상대효율이득은 평균 52.74%의 효율이득이 발생하였다. 기존 직접추정값 대비 기존 복합추정값들의 상대효율이득은 평균 57.60%의 효율이득이 발생하였다. 기존 직접추정값 대비 기존 Multi-level 모형을 이용한 계층적 베이즈 추정값들의 상대효율이득은 평균 5.89%, 기존 시계열 및 횡단면 모형을 이용한 추정에서는 평균 2.11%의 상대효율이득이 발생하였다.

5월 조사결과와 마찬가지로 6월 조사 결과에서도 복합추정법이 추정값들의 안정성 및 효율측면에서 탁월한 결과를 보인다. 현행 조사구를 이용하여 추정된 합성 추정값과 복합 추정값들에 대한 상대효율이득이 표본 조사구 수 증편 후의 직접 추정값들에 대한 평균 상대효율이득인 26.64%보다 훨씬 크게 나타난다. 특히 복합 추정량은 다른 추정량들에 비해 월등한 효율을 보인다.

<표5.25> 직접추정량 대비 상대효율 이득 비교(2001년 6월)

시도	시군구	직접추정량 (기존조사구 이용)		직접추정량 (증편조사구 이용)			합성추정량 (기존조사구 이용)			복합추정량 (기존조사구 이용)		
		추정값	RSE	추정값	RSE	효율 이득 (%)	추정값	RSE	효율 이득 (%)	추정값	RSE	효율 이득 (%)
광주광역시	동구	2,331	25.83	3,465	18.36	28.92	2,685	12.67	50.95	2,854	11.40	55.87
	서구	1,819	68.77	6,612	18.27	73.43	6,209	12.87	81.29	5,345	13.82	79.90
	남구	13,336	24.93	6,597	21.69	13.00	5,054	12.84	48.50	6,005	11.85	52.47
	북구	8,986	16.65	9,128	16.65	0.00	10,257	12.82	23.00	10,508	10.21	38.68
	광산구	2,533	32.21	3,203	26.46	17.85	4,801	12.89	59.98	4,293	12.53	61.10
	합계	29,005		29,005		(26.64)	29,005		(52.74)	29,005		(57.60)

시도	시군구	Multi-level 모형 추정량 (기존조사구 이용)			시계열 및 횡단면 모형 추정량 (기존조사구 이용)		
		추정값	RSE	효율 이득 (%)	추정값	RSE	효율 이득 (%)
광주광역시	동구	3,458	450.1	32.94	3,406	797.6	-1.69
	서구	4,943	1311.0	48.69	5,848	1305.0	63.61
	남구	5,501	2878.0	-179.31	6,074	1356.0	-0.45
	북구	11,955	118.5	92.08	10,157	1946.0	-29.07
	광산구	3,148	495.2	35.03	3,520	1232.0	-21.83
	합계	29,005		(5.89)	29,005		(2.11)

(나) 충청북도

2001년 5월과 6월 충청북도 경제활동인구조사에 대한 설계기반 소지역 추정량들의 상대효율 이득에 관련한 결과가 다음 <표5.26>과 <표5.27>에 주어졌다.

여기에서 상대효율 이득은 표본 조사구 증편에 따른 상대표준오차의 절감량을 백분율로 나타낸 값이며 다음 식과 같이 주어진다.

$$\text{상대효율 이득} = \frac{RSE(\text{기존}) - RSE(\text{증편})}{RSE(\text{기존})} \times 100\ (\%)$$

다음 <표5.26>은 5월의 충청북도 경제활동인구조사 결과이다. 기존 직접추정값 대비 증편조사 직접추정값들의 상대효율이득, 기존 합성추정값 대비 증편조사 합성추정값들의 상대효율이득과 기존 복합추정값 대비 증편조사 복합추정값들의 상대효율이득을 요약하였다.

5월 충청북도의 시군단위 소지역들에 대한 기존 직접추정값 대비 증편조사 직접추정값들의 상대효율이득은 최소 -8.437%에서 최대 65.48%의 범위에 있고 평균 24.49%의 효율이득이 발생하였다.

충주시와 음성군의 경우에는 표본 조사구 증편 전 보다 오히려 효율이 감소하였다. 기존 합성추정값 대비 증편조사 합성추정값들의 상대효율이득은 평균 36.94%, 증편조사 복합추정값들의 상대효율이득은 평균 36.56%의 효율이득이 발생하였다.

<표5.26> 설계기반 추정량들의 상대효율 이득(2001년 5월)

시도	시군구	직접추정량			합성추정량			복합추정량		
		RSE (기존 조사)	RSE (증편 조사)	효율 이득 (%)	RSE (기존 조사)	RSE (증편 조사)	효율 이득 (%)	RSE (기존 조사)	RSE (증편 조사)	효율 이득 (%)
충청북도	청주시	19.62	19.58	0.20	11.51	10.42	9.47	9.94	9.30	6.43
	충주시	24.88	**26.77**	-7.60	11.93	10.74	9.97	10.74	9.99	6.98
	제천시	**63.25**	**26.64**	57.88	11.97	10.68	10.78	12.29	9.94	19.12
	보은군	**86.31**	23.05	73.29	**29.28**	15.55	46.89	**27.76**	13.83	50.18
	옥천군	**42.15**	25.25	40.09	**30.53**	16.19	46.97	24.96	14.04	43.75
	영동군	**75.00**	**55.85**	25.53	**30.48**	16.41	46.16	**30.37**	16.12	46.92
	괴산군	**83.31**	**65.48**	21.40	**29.43**	15.47	47.43	**28.34**	15.30	46.01
	음성군	**43.28**	**46.93**	-8.43	**29.34**	15.42	47.44	24.56	14.64	40.39
	청원군	**40.03**	**32.79**	18.09	**30.39**	16.19	46.73	25.15	14.56	42.11
	진천군	-	**48.47**	-	**29.82**	15.70	47.35	**29.82**	14.89	50.07
	단양군	-	**46.53**	-	**29.57**	15.63	47.14	**29.57**	14.72	50.22
	평균			24.49			36.94			36.56

6월의 경제활동인구조사에서 표본 조사구 수 추가에 따른 각 추정값들의 상대효율이득은 다음 <표5.27>에 주어졌다.

6월 충청북도의 시군단위 소지역들에 대한 기존 직접추정값 대비 증편조사 직접 추정값들의 상대효율이득은 최소 -4.84%에서 최대 56.68%의 범위에 있고 평균 30.41%의 효율이득이 발생하였다. 충주시의 경우에는 표본 조사구 증편 전 보다 오히려 효율이 감소하였다.

기존 합성추정값 대비 증편조사 합성추정값들의 상대효율이득은 평균 29.77%, 증편조사복합 추정값들의 상대효율이득은 평균 29.04%의 효율이득이 발생하였다.

<표5.27> 설계기반 추정량들의 상대효율 이득(2001년 6월)

시도	시군구	직접추정량			합성추정량			복합추정량		
		RSE (기존 조사)	RSE (증 편 조사)	효율 이득 (%)	RSE (기존 조사)	RSE (증편 조사)	효율 이득 (%)	RSE (기존 조사)	RSE (증편 조사)	효율 이득 (%)
충청북도	청주시	22.80	22.80	0.00	14.99	13.51	9.87	12.53	11.75	6.23
	충주시	**26.02**	**27.28**	-4.84	15.10	13.54	10.33	13.05	12.18	6.67
	제천시	**72.83**	**31.55**	56.68	15.19	13.56	10.73	15.17	12.48	17.73
	보은군	-	**51.42**	-	25.31	16.36	35.36	25.31	15.91	37.14
	옥천군	**78.71**	**32.09**	59.23	**25.86**	17.48	32.41	25.12	15.50	38.30
	영동군	**75.20**	**67.82**	13.80	**25.81**	17.57	31.93	25.06	17.19	31.40
	괴산군	**117.9**	**64.41**	45.37	25.32	16.67	34.16	24.79	16.43	33.72
	음성군	**36.54**	**27.17**	25.64	25.35	16.51	34.87	22.14	14.07	36.45
	청원군	**71.35**	**37.51**	47.43	**25.95**	17.12	34.03	**25.57**	15.56	39.15
	진천군	-	**38.90**	-	**25.61**	16.55	35.38	**25.61**	15.32	40.18
	단양군	-	**103.40**	-	**39.88**	16.60	58.38	25.27	17.05	32.52
	평균			30.41			29.77			29.04

다음 <표5.28>과 <표5.29>에서는 기존 직접추정값 대비 증편조사 직접추정값들의 상대효율이득과 기존 직접추정값 대비 기존 소지역 추정값들의

상대효율이득을 요약하였다. 기존 직접추정값 대비 기타 소지역 추정값들의 상대효율 이득은 다음 식으로 주어진다.

$$\text{상대효율 이득} = \frac{\text{직접추정값의 } RSE - \text{기타 추정값의 } RSE}{\text{직접추정값의 } RSE} \times 100\,(\%)$$

다음 <표5.28>은 2001년 5월의 충청북도 경제활동인구조사 자료를 이용한 추정결과이다.

기존 직접추정값 대비 표본 조사구 수 증편 후의 직접 추정값들의 상대효율이득은 최소 -8.43%에서 최대 73.29%의 범위에 있고, 표본 조사구 수 증편 전에 비해 평균 24.49%의 효율이득이 발생하였다. 오히려 충주시와 음성군의 경우는 표본조사구 증편 후 각각 -7.60%와 -8.43%의 효율의 손실이 발생하였다.

기존 직접추정값 대비 기존 합성추정값들의 상대효율이득은 평균 49.82%의 효율이득이 발생하였고, 기존 직접추정값 대비 기존 복합추정값들에 대한 상대효율이득은 평균 55.69%의 효율이득이 발생하였다.

기존 직접추정값 대비 Multi-level 모형을 이용한 계층적 베이즈 추정값들의 상대효율이득은 평균 47.86%, 기존 직접추정값 대비 시계열 및 횡단면 모형을 이용한 계층적 베이즈 추정값들의 상대효율이득은 평균 35.22%의 효율이득이 발생하였다.

증편조사 직접추정값들의 평균 상대효율이득인 24.49%보다 기존조사구를 이용하여 추정된 합성추정값, 복합추정값과 모형기반 추정값들의 상대효율이득이 훨씬 크다. 특히 복합추정값들의 상대효율이득이 다른 추정값들의 상대효율이득보다 크게 나타나며 매우 안정적이다.

<표5.28> 기존 직접추정값 대비 상대효율 이득 비교(2001년 5월)

시도	시군구	직접추정량 (기존조사구 이용)		직접추정량 (증편조사구 이용)			합성추정량 (기존조사구 이용)			복합추정량 (기존조사구 이용)		
		추정값	RSE	추정값	RSE	효율 이득 (%)	추정값	RSE	효율 이득 (%)	추정값	RSE	효율 이득 (%)
충청북도	청주시	9,929	19.62	8,357	19.58	0.20	9,341	11.51	41.34	9,434	9.94	49.34
	충주시	3,296	24.88	3,084	26.77	-7.60	2,837	11.93	52.05	2,939	10.74	56.83
	제천시	751	**63.25**	2,535	26.64	57.88	1,798	11.97	81.08	1,604	12.29	80.57
	보은군	263	**86.31**	749	23.05	73.29	300	29.28	66.08	322	27.76	67.84
	옥천군	790	**42.15**	837	25.25	40.09	576	30.53	27.57	678	24.96	40.78
	영동군	252	**75.00**	250	**55.84**	25.55	577	30.48	59.36	461	30.37	59.51
	괴산군	779	**83.31**	503	**65.48**	21.40	295	29.43	64.67	331	28.34	65.98
	음성군	1,294	**43.28**	661	**46.93**	-8.43	882	29.34	32.21	1,040	24.56	43.25
	청원군	767	**40.03**	918	**32.79**	18.09	1,343	30.39	24.08	1,054	25.15	37.17
	진천군	779	-	584	**48.47**	-	610	29.82	-	665	29.82	-
	단양군	0	-	422	**46.53**	-	341	29.57	-	372	29.57	-
	합계	18,900		18,900		(24.49)	18,900		(49.82)	18,900		(55.69)

시도	시군구	Multi-level 모형 추정량 (기존조사구 이용)			시계열 및 횡단면 모형 추정량 (기존조사구 이용)		
		추정값	RSE	효율 이득 (%)	추정값	RSE	효율 이득 (%)
충청북도	청주시	10,196	8.04	59.03	10,150	15.03	23.36
	충주시	3,093	31.52	-26.71	2,937	22.90	7.94
	제천시	687	21.94	65.31	888	**96.90**	-53.21
	보은군	298	23.59	72.67	303	22.73	73.66
	옥천군	1,052	**29.19**	30.74	1,187	11.38	72.99
	영동군	251	26.08	65.22	221	54.19	27.75
	괴산군	794	**15.34**	81.58	790	17.66	78.81
	음성군	1,109	16.28	62.39	986	15.98	63.07
	청원군	559	31.81	20.53	589	30.98	22.61
	진천군	402	**46.16**	-	483	26.61	-
	단양군	459	29.89	-	366	24.31	-
	합계	18,900		(47.86)	18,900		(35.22)

<표5.29>는 2001년 6월 충청북도 소지역 추정결과이다.

<표5.29> 기존 직접추정값 대비 상대효율 이득 비교(2001년 6월)

시도	시군구	직접추정량 (기존조사구 이용)		직접추정량 (중편조사구 이용)			합성추정량 (기존조사구 이용)			복합추정량 (기존조사구 이용)		
		추정값	RSE	추정값	RSE	효율 이득 (%)	추정값	RSE	효율 이득 (%)	추정값	RSE	효율 이득 (%)
충청북도	청주시	10,549	22.80	8,985	22.80	0.00	9,981	14.99	34.25	10,021	12.53	45.04
	충주시	3,297	**26.02**	3,112	**27.28**	-4.84	3,041	15.10	41.97	3,088	13.05	49.85
	제천시	1,108	**72.83**	2,857	**31.55**	56.68	1,932	15.19	79.14	1,844	15.17	79.17
	보은군	0	-	425	**51.42**	-	226	25.31	-	235	25.31	-
	옥천군	263	**78.71**	498	**32.09**	59.23	436	**25.86**	67.15	404	25.12	68.09
	영동군	250	**75.20**	281	**67.82**	9.81	437	**25.81**	65.68	393	25.06	66.68
	괴산군	263	**117.9**	501	**64.41**	45.37	223	25.32	78.52	232	24.79	78.97
	음성군	1,828	**36.54**	641	**27.17**	25.64	667	25.35	30.62	767	22.14	39.41
	청원군	513	**71.35**	714	**37.51**	47.43	1,017	**25.95**	63.63	851	**25.57**	64.16
	진천군	513	-	498	**38.90**	-	462	**25.61**	-	480	**25.61**	-
	단양군	0	-	72	**103.4**	-	162	**39.88**	-	266	25.27	-
	합계	18,584		18,584		(29.92)	18,584		(57.62)	18,584		(61.42)

시도	시군구	Multi-level 모형 추정량 (기존조사구 이용)			시계열 및 횡단면 모형 추정량 (기존조사구 이용)		
		추정값	RSE	효율 이득(%)	추정값	RSE	효율 이득(%)
충청북도	청주시	11,071	12.73	44.12	10,799	5.97	73.83
	충주시	3,205	3.57	86.27	3,370	7.68	70.47
	제천시	678	**42.18**	42.08	786	**64.21**	11.84
	보은군	242	**38.11**	-	270	**31.96**	-
	옥천군	513	**65.37**	16.95	719	21.18	73.09
	영동군	259	13.32	82.28	185	**85.97**	-14.32
	괴산군	427	**44.69**	62.09	389	**38.12**	67.66
	음성군	913	**53.91**	-47.51	898	20.02	45.22
	청원군	579	**34.88**	51.12	362	**50.31**	29.48
	진천군	440	**46.34**	-	487	**35.78**	-
	단양군	257	**57.53**	-	319	**32.79**	-
	합계	18,584		(42.18)	18,584		(44.66)

6월의 충청북도의 경제활동인구조사에서 기존 직접추정값 대비 증편조사 직접추정값들의 상대효율이득은 최소 -4.84%에서 최대 59.23%의 범위에 있고, 표본 조사구 수 증편 전에 비해 평균 29.92%의 효율이득이 발생하였다. 충주시의 경우에는 증편 전보다 오히려 -4.84%의 효율손실이 발생하였다.

기존 직접추정값 대비 기존 합성추정값들의 상대효율이득은 평균 57.62%의 효율이득이 발생하였고, 기존 직접추정값 대비 기존 복합추정값들에 대한 상대효율이득은 평균 61.42%의 효율이득이 발생하였다.

기존 직접추정값 대비 기존 Multi-level 모형을 이용한 계층적 베이즈 추정값들의 상대효율이득은 평균 42.18%, 기존 시계열 및 횡단면 모형을 이용한 추정에서는 평균 44.66%의 상대효율이득이 발생하였다.

현행 조사구를 이용하여 추정된 합성 추정값과 복합 추정값들에 대한 평균 상대효율이득은 각각 57.62%와 61.42%로써 표본 조사구 수 증편 후의 직접 추정값들에 대한 평균 상대효율이득 29.92%보다 두 배 정도 크게 나타났다. 특히 복합 추정값들은 다른 설계기반 추정값 및 모형기반 추정값들에 비해 월등한 효율을 보인다.

<표5.24>, <표5.25>, <표5.28>와 <표5.28>의 결과로부터 다음과 같은 사실을 확인할 수 있다.

첫째, 표본조사구 수를 다수 증편하더라도 직접추정값들의 효율이득은 현저히 좋아지지는 않는다.

둘째, 표본조사구 증편 후의 직접추정값들의 효율이득보다 현행 경제활동인구조사에서 추정된 합성추정값 또는 복합추정값들의 효율이득이 훨씬 높

게 나타난다.

셋째, 모형기반 추정량들의 평균 상대효율이득이 합성추정량이나 복합추정량보다 작게 나타나고 몇몇 소지역에 대한 모형기반 추정값들의 변동이 심하게 나타나는 것은 월별 직접추정값들의 변동이 심하다는 사실을 보여주는 결과이다.

따라서 모형기반 추정값들은 소지역에 대한 표본조사구 증편 시 참고자료로 활용될 수 있다.

경제활동인구조사에서 추정치 산정에 문제가 있는 몇몇 시군구 단위 소지역들에 대한 추가 조사구 산정 시 모형기반 추정값들을 참조한 후 추가 조사구 수를 결정하고, 시군구 단위 소지역들에 대한 추정에서는 설계기반 추정법인 복합 추정법을 활용한다면 모든 시군구 단위 소지역 추정값들에 대한 상대표준오차의 목표 요구정도를 확보할 수 있을 것으로 판단된다.

5.4 프로그램 알고리즘

Multi-level 모형과 시계열 및 횡단면 모형을 이용한 계층적 베이즈 추정량은 전문 프로그램인 WinBUGS 프로그램을 이용하여 추정되며, 추정 과정 중 사전분포의 선정 및 초기값 배정에서 전문가의 경험적인 능력이 요구되므로 여기에서는 언급을 하지 않는다.

직접 추정값, 합성 추정값, 복합 추정값 및 추정값들의 분산을 계산하기 위한 개략적인 프로그램 알고리즘을 단계별로 설명하기로 한다. 프로그램의 전반적인 흐름도 및 세부적인 소스코드에 대한 설명은 부록에 상세하게 수록하였으므로 참고하기 바란다.

5.4.1 직접추정값과 추정값의 분산계산

(Step1) 직접추정값 계산

대영역 내의 I개의 소지역에 대한 실업자 총계 추정값은 경제활동인구 조사 자료에 근거하여 다음 직접 추정 공식을 이용하여 계산한다.

$$\hat{Y}_{i\cdot} = \sum_{s} {}_{s}\hat{Y}_{i\cdot} \quad , \quad i=1,2,\cdots,I;\ s=1,2;\ h=1,2,\cdots,n_i$$

$$= \sum_{s}\sum_{h} {}_{s}\hat{Y}_{ih}$$

$$= \sum_{s}\sum_{h} {}_{s}M_i \ {}_{s}Y_{ih} \ ,$$

여기에서 s= 남, 녀(1, 2),

n_i= 소지역 i의 조사구 수,

${}_{s}Y_{ih}$ = 소지역 i의 실업자 수(남, 녀)

${}_{s}M_i = {}_{s}\hat{X}_{i\cdot} / {}_{s}X_{i\cdot}$; 승수(주어진 값),

${}_{s}\hat{X}_{i\cdot}$ = 소지역 i의 15세 이상의 상주추정인구(남, 녀),

${}_{s}X_{i\cdot}$ = 경제활동인구조사에서 15세 이상의 인구.

(Step2) 직접추정값의 분산 계산

i번째 소지역에 대한 직접 추정값의 분산 공식은

$$Var(\hat{Y}_{i\cdot}) = \sum_{s=1}^{2} Var({}_{s}\hat{Y}_{i\cdot}) + 2\,Cov({}_{1}\hat{Y}_{i\cdot}\ ,\ {}_{2}\hat{Y}_{i\cdot})$$

$$= \sum_{s=1}^{2} {}_{s}M_i^2\, Var\Big(\sum_{h=1}^{n_i} {}_{s}Y_{ih}\Big) + 2\,{}_{1}M_i\ {}_{2}M_i Cov\Big(\sum_{h=1}^{n_i} {}_{1}Y_{ih}\ ,\ \sum_{h=1}^{n_i} {}_{2}Y_{ih}\Big)$$ 을 적용

하고, 이를 이용한 분산 추정공식은 다음 식을 적용하여 계산한다.

$$\widehat{Var}(\hat{Y}_{i\cdot}) = \sum_{s=1}^{2} {}_{s}M_i^2 \left(\xi_i \sum_{k=1}^{n_i} {}_{s}U_{ik}^2\right) + 2\,{}_{1}M_i\,{}_{2}M_i \left(\xi_i \sum_{k=1}^{n_i} {}_{1}U_{ik} \cdot {}_{2}U_{ik}\right) \quad ,$$

여기에서 $${}_{s}U_{ik} = d_s Y_{ik} - {}_{s}\rho_i \cdot d_s X_{ik} ,$$

$$d_s Y_{ik} = {}_{s}Y_{ik} - {}_{s}Y_{i,k+1} ,$$

$$d_s X_{ik} = {}_{s}X_{ik} - {}_{s}X_{i,k+1} ,$$

$${}_{s}\rho_i = {}_{s}Y_{i\cdot} / {}_{s}X_{i\cdot} ,$$

$$\xi_i = [1 - n_i/(10N_i)]n_i/[2(n_i - 1)] ,$$

N_i=소지역 i의 모집단 조사구수.

5.4.2 합성추정값, 복합추정값과 추정값들의 분산계산

설계 기반 추정값으로써 합성추정값과 복합추정값 및 추정값들의 추정 분산에 대한 계산 절차를 소개하기 위해 대영역인 충청북도 내의 시군 단위의 소지역들에 대해서 언급한다.

(Step1) 2001년 5월의 조사자료에서 충청북도를 크게 시지역과 군지역으로 2개의 그룹으로 구분하여 각 그룹에 대한 경제활동인구와 실업자수를 계산한다.

① 시지역의 남자경제활동인구수, 실업자수 계산

② 시지역의 여자경제활동인구수, 실업자수 계산

③ 군지역의 남자경제활동인구수, 실업자수 계산

④ 군지역의 여자경제활동인구수, 실업자수 계산

(Step2) 5월의 조사자료에서 시(군)지역의 각 성별(남, 여)에 대한 경제활동참가율을 계산(K1～K4)

① 시지역 남자의 경제활동참가율(K1)

= 시지역의 남자 경활인구÷시지역 남자 15세 이상 인구

② 시지역 여자의 경제활동참가율(K2)

= 시지역 여자 경활인구÷시지역 여자 15세 이상 인구

③ 군지역 남자의 경제활동참가율(K3)

= 군지역 남자 경활인구÷군지역 남자 15세 이상 인구

④ 군지역 여자의 경제활동참가율(K4)

= 군지역 여자 경활인구÷군지역 여자 15세 이상 인구

(Step3) 5월의 조사자료에서 시(군)지역의 남,녀의 범주별 실업률계산

① 시지역의 범주별 실업률

o 시지역의 남자 15-34의 실업률(SU1)

=시지역 남자15-34의 실업자수÷시지역 남자15-34의 경활인구

o 시지역의 남자 35이상의 실업률(SU2)

=시지역 남자35이상의 실업자수÷시지역 남자35이상의 경활인구

o 시지역의 여자 15-34의 실업률(SU3)

=시지역 여자15-34의 실업자수÷시지역 여자15-34의 경활인구

o 시지역의 여자 35이상의 실업률(SU4)

=시지역 여자35이상의 실업자수÷시지역 여자35이상의 경활인구

② 군지역의 범주별 실업률

o 군지역의 남자 15-34의 실업률(GU1)

=군지역 남자15-34의 실업자수÷군지역 남자15-34의 경활인구

o 시지역의 남자 35이상의 실업률(GU2)

=군지역 남자35이상의 실업자수÷군지역 남자35이상의 경활인구

o 시지역의 여자 15-34의 실업률(GU3)

=군지역 여자15-34의 실업자수÷군지역 여자15-34의 경활인구

o 시지역의 여자 35이상의 실업률(GU4)

=군지역 여자35이상의 실업자수÷군지역 여자35이상의 경활인구

(Step4) 5월의 조사자료에서 시지역과 군지역을 각각 4개의 범주(남15-34세, 남35세이상, 여15-34세, 여35세이상)로 구분하여 경제활동인구 계산

① 시지역의 범주별 경제활동인구수

o 시지역 전체(청주+충주+제천)에서 남15-34의 경제활동인구수(S1)

o 시지역 전체에서 남35이상의 경제활동인구수(S2)

o 시지역 전체에서 여15-34의 경제활동인구수(S3)

o 시지역 전체에서 여35이상의 경제활동인구수(S4)

② 군지역의 범주별 경제활동인구수

o 군지역 전체에서 남15-34의 경제활동인구수(G1)

o 군지역 전체에서 남35이상의 경제활동인구수(G2)

o 군지역 전체에서 여15-34의 경제활동인구수(G3)

o 군지역 전체에서 여35이상의 경제활동인구수(G4)

(Step5) 통계청 자료에서 2001년 5월의 소지역의 추계인구 데이터를 입력한다.

(예시) 통계청으로부터 주어지는 '01년 5월 추계인구

범주	청주	충주	제천	보은	옥천	영동	괴산	음성	청원	진천	단양
남15-34	109786	35407	21830	4442	8713	9012	4468	12493	19260	8328	4926
남35이상	114522	47953	33288	11336	15183	14714	11786	20879	28513	13503	10077
여15-34	115760	31081	19704	3506	3839	7071	3316	4142	17233	7611	4163
여35이상	118750	51614	36025	13194	17408	17719	13758	21424	31344	14425	11015

(Step6) (Step5)의 각 셀에 대한 경제활동인구 계산

(E1(1)~E1(4), E2(1)~E2(4), …, E11(1)~E11(4))

① 시지역(청주시의 경우)

o 남15-34의 경제활동인구(E1(1))

= 109786*시지역 남자의 경제활동참가율(K1)

o 남35이상의 경제활동인구(E1(2))

=114522*K1

o 여15-34의 경제활동인구(E1(3))

=115760*시지역 여자의 경제활동참가율(K2)

o 여35이상의 경제활동인구(E1(4))

=118750*K2

② 군지역(보은군의 경우)

o 남15-34의 경제활동인구(E4(1))

=4442*군지역 남자의 경제활동참가율(K3)

o 남34이상의 경제활동인구(E4(2))

=11336*K3

o 여15-34의 경제활동인구(E4(3))

=3506*군지역 여자의 경제활동참가율(K4)

o 여35이상의 경제활동인구(E4(4))

=13194*K4

(Step7) 각 소지역에 대한 합성추정값의 분산계산을 계산한다.

① 시지역 합성분산(청주시의 경우)

$$= \left(\frac{E1(1)}{S1}\right)^2 \left(M_1^2\zeta_1 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{i1h}^2\right)$$

$$+ \left(\frac{E1(2)}{S2}\right)^2 \left(M_2^2\zeta_2 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{i2h}^2\right)$$

$$+ \left(\frac{E1(3)}{S3}\right)^2 \left(M_3^2\zeta_3 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{i3h}^2\right)$$

$$+ \left(\frac{E1(4)}{S4}\right)^2 \left(M_4^2\zeta_4 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{i4h}^2\right)$$

$$+ 2*\Bigg[\left\{\frac{E1(1)*E1(2)}{S1*S2}\left(M_1M_2\zeta_1 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{i1h}U_{i2h}\right)\right\}$$

$$+ \left\{\frac{E1(1)*E1(3)}{S1*S3}\left(M_1M_3\zeta_1 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{i1h}U_{i3h}\right)\right\}$$

$$+ \left\{\frac{E1(1)*E1(4)}{S1*S4}\left(M_1M_4\zeta_1 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{i1h}U_{i4h}\right)\right\}$$

$$+ \left\{\frac{E1(2)*E1(3)}{S2*S3}\left(M_2M_3\zeta_2 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{i2h}U_{i3h}\right)\right\}$$

$$+ \left\{\frac{E1(2)*E1(4)}{S2*S4}\left(M_2M_4\zeta_2 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{i2h}U_{i4h}\right)\right\}$$

$$+ \left\{\frac{E1(3)*E1(4)}{S3*S4}\left(M_3M_4\zeta_3 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{i3h}U_{i4h}\right)\right\} \Bigg],$$

여기에서 $M_1 = M_2 = M_3 = M_4$: 충북 시지역의 승수,

$$\zeta_j = \frac{n_j}{2(n_j-1)}, \quad n_j : j\text{번째 범주의 조사구수}$$

② 군지역 합성분산(보은군의 경우)

$$= \left(\frac{EA(1)}{G1}\right)^2 \left(M_1^2\zeta_1 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{1ih}^2\right)$$

$$+ \left(\frac{EA(2)}{G2}\right)^2 \left(M_2^2\zeta_2 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{2ih}^2\right)$$

$$+ \left(\frac{EA(3)}{G3}\right)^2 \left(M_3^2\zeta_3 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{3ih}^2\right)$$

$$+ \left(\frac{EA(4)}{G4}\right)^2 \left(M_4^2\zeta_4 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{4ih}^2\right)$$

$$+ 2*\left[\left\{\frac{EA(1)*EA(2)}{G1*G2}\left(M_1M_2\zeta_1 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{1ih}U_{2ih}\right)\right\}\right.$$

$$+ \left\{\frac{EA(1)*EA(3)}{G1*G3}\left(M_1M_3\zeta_1 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{1ih}U_{3ih}\right)\right\}$$

$$+ \left\{\frac{EA(1)*EA(4)}{G1*G4}\left(M_1M_4\zeta_1 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{1ih}U_{4ih}\right)\right\}$$

$$+ \left\{\frac{EA(2)*EA(3)}{G2*G3}\left(M_2M_3\zeta_2 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{2ih}U_{3ih}\right)\right\}$$

$$+ \left\{\frac{EA(2)*EA(4)}{G2*G4}\left(M_2M_4\zeta_2 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{2ih}U_{4ih}\right)\right\}$$

$$+ \left.\left\{\frac{EA(3)*EA(4)}{G3*G4}\left(M_3M_4\zeta_3 \sum_{i=1}^{I}\sum_{h=1}^{n_h} U_{3ih}U_{4ih}\right)\right\}\right] ,$$

여기에서 $M_1 = M_2 = M_3 = M_4$: 충북 군지역의 승수,

$$\zeta_j = \frac{n_j}{2(n_j-1)}, \quad n_j : j\text{번째 범주의 조사구수}$$

(Step8) 각 소지역의 가중값을 계산한다.

① 청주시의 가중치(ω_1)

$$w_1 = \frac{\text{청주시합성분산}}{\text{청주시합성분산} + \text{청주시직접추정분산}}$$

② 보은군의 가중치(ω_4)

$$w_4 = \frac{\text{보은군합성분산}}{\text{보은군합성분산} + \text{보은군직접추정분산}}$$

(Step9) 각 소지역에 대한 합성추정값을 계산한다.

① 청주시 합성추정값

= {E1(1)*SU1}+{E1(2)*SU2}+{E1(3)*SU3}+{E1(4)*SU4}

② 보은군 합성추정값

= {E4(1)*GU1}+{E4(2)*GU2}+{E4(3)*GU3}+{E4(4)*GU4}

(Step10) 각 소지역에 대한 복합추정값을 계산한다.

① 청주시의 복합추정값

= w_1*청주시직접추정값 + (1- w_1)*청주시합성추정값

② 보은군의 복합추정값

= w_4*보은군직접추정값 + (1- w_4)*보은군합성추정값

(Step11) 각 소지역의 복합추정값의 표준오차와 CV(=RSE)값을 계산한다.

① 표준오차(S.E) 계산

o 청주시의 복합추정값의 표준오차

= sqrt(w_1^2*청주시직접추정분산 + $(1-w_1)^2$*청주시합성분산)

o 보은군의 복합추정값의 표준오차

= sqrt(w_4^2*보은군직접추정분산 + $(1-w_4)^2$*보은군합성분산)

② CV(=RSE)값 계산

o 청주시의 CV값

= 청주시의 복합추정값의 표준오차÷청주시의 복합추정값

o 보은군의 CV값

= 보은군의 복합추정값의 표준오차÷보은군의 복합추정값

(Step) 각 소지역에서의 추정값, 표분오차와 CV값을 출력한다.

제 6장 시군구 실업통계의 표본설계와 추정모형 개발

6.1 개 요

경제활동인구조사의 표본설계는 월별 경제활동인구의 실업자와 취업자를 현재 사용 중인 전국 단위와 시도 단위까지 추계할 목적으로 연구되었으나 매월 통계청에서 발표하는 경제활동인구통계의 종류는 경제활동 총괄(시도와 전국단위), 전국 단위의 연령계층별 경제활동인구, 취업자(교육정도별, 산업별, 직업별, 종사상 지위별, 취업시간대별, 연령계층별 실업자 및 실업률, 교육정도별 실업자 및 실업률과 지역별 경제활동인구 총괄 등이 있다.

표본설계 당시에는 시도 단위와 전국 단위의 경제활동인구(취업자, 실업자 포함)의 통계만을 일정 수준의 신뢰도와 정확도를 갖추도록 표본크기를 결정하고 표본을 배분하였으나 연령계층별이나 교육정도별 또는 산업별 등의 보다 더 세분화된 범주에 대한 통계를 생산해야 할 경우에는 이들에 대한 추정오차는 사전에 통제될 수 없을 것이다.

또한 앞에서 언급했듯이 시군구 단위의 경제활동인구(취업자, 실업자 등)를 추정하고자 할 때에는 추정오차가 커지기 때문에 정부의 공식통계로 사용하는데 문제가 있을 수 있으므로 표본설계 시에는 현재 당면하고 있는 실태뿐만 아니라 앞으로의 상황까지 예견하여 정부 공식통계의 활용성이 최대화 되도록 다양한 분야의 전문가들의 의견을 수렴해야 할 것이다.

본 장에서는 2002년도 경제활동인구조사의 표본설계에서 시군구의 실업통계 생산을 대 전제로 했을 때 표본설계 시에 고려되어야 할 요소를 알아보고 적합한 설계절차를 제안하고 이에 따라 시군구의 실업통계를 일정한 정확도를 갖출 수 있는 소지역 추정법을 개발하고 이에 대한 알고리즘을 연

구하는 방향을 제시하고자 한다.

6.2 표본 설계

5장에서 소지역 추정법을 이용한 시군구 실업통계 작성의 가능성과 타당성을 실증조사를 통해 살펴보았다. 소지역 추정법 중에서도 복합 추정법의 사용을 전제로 하여 표본설계의 연구방향을 언급하겠다.

표본설계에서 우선적으로 연구해야 할 사항은 모집단 분석이다. 2002년 인구주택 총 조사의 10% 표본조사구에 대해서 복합 추정법을 적용하는데 필요한 모집단의 구조분석 및 파악과 실업률을 잘 표현할 수 있는 지역적인 구분에 대한 분석이 필수적이다.

지금까지는 모집단을 광역시와 도별로 구분하고 그 내부에서 동부와 읍면부로 나누어서 25개 층으로 분할하여 표본을 추출하고 승수를 산정하여 추정에 이용하였으나, 광역시와 도지역으로 나누고 도지역에서는 산업 특성이나 지리적인 특성을 감안하여 권역별로 구분하고 권역 내에서도 실업의 구조적 특성을 나타낼 수 있도록 도시지역, 준 산업지역과 농림업 지역으로 3개로 구분하여 도지역을 13개 층으로 분할하여 전국이 20개 층이 되도록 층화하는 방안을 연구해야 할 것이다.

층화를 전제로 한 모집단 분석은 행정구역 경계를 떠나서 성별(남, 여)-연령대(15-34세, 35세 이상)별로 실업률이 유사한 인접 시군구를 묶어서 층을 만들 수 있어야 할 것이며 이와 같은 연구는 지리적 특성과 인구주택 총조사 자료를 이용하여 시행 착오적이고 반복적인 컴퓨터 계산을 통해서 수행되어야 할 것이다.

모집단의 층화가 이루어진 후에는 실업통계의 분석 및 생산단위의 결정과 추정값의 목표정도를 결정하는 것이다. 추정값의 목표정도와 실업통계 발표단위의 결정은 통계청의 조사인력과 예산 범위 내에서 이루어져야 한다. 일반적으로 통계청의 조사인력과 예산범위가 정해지면 표본조사구의 수와 표본조사 가구의 수가 자동으로 정해지므로 추정값의 추정오차를 최소화하는 최선의 추정량을 찾아내는 것이다

이와 같이 표본설계의 연구내용들이 서로 연계되어 있어서 독립적으로 최선 방안을 설명하는 것은 불가능할 수 있으나 표본설계에서 빠뜨려서는 안될 사항들을 설명하여 앞으로 연구 진행에서 참고가 되도록 하겠다.

6.2.1 분석단위 결정

시군구 단위의 실업통계 작성을 전제로 표본설계를 연구할지라도 시군구별로 인구수와 면적의 규모에서 많은 차이가 있고 또한 230여 시군구별로 매월 실업통계를 생산해서 발표할 수 있도록 조사인력이나 예산의 지원이 가능할 것인지를 고려해야 할 것이다. 이는 정책적으로 결정될 사항이지만 통계 생산의 활용성과 경제성도 함께 고려되어야 할 것이다.

예를 들어 충청북도 시군구의 인구 규모를 2000년 말 주민등록인구수 기준으로 살펴보면 청주시 58만명, 충주시 22만명, 청원군 12만명 이지만 단양군은 4만명, 보은군은 4만 3천명으로 모든 시군구에 대해서 획일적으로 매월 실업통계를 생산하여 발표하는 것보다는 일정 규모 이상의 시군구에 대해서는 매월 실업통계를 발표하지만 소규모의 시군구에 대해서는 분기별로 실업통계를 발표하는 방안도 함께 연구 검토되어야 할 것이다.

표본설계를 기획할 시점에서 분석단위를 결정했을 지라도 사회 환경 변화 또는 정책 입안 시 고려되지 않았던 분석 단위의 통계를 생산해야 할 경우에는 일정 수준의 신뢰수준을 갖추지 못하더라도 정책 결정이나 입안의 참고자료로 활용할 수 있도록 새로운 분석단위에 대해서 통계를 작성할 수 있는 소지역 추정법이 고려되어야 할 것이다.

6.2.2 표본 조사구의 크기 결정과 배분

표본설계에서 표본추출법과 추정방법이 주어지고 목표정도가 결정되면 표본의 크기는 산정할 수 있으나 경제활동인구조사와 같이 대규모이면서 일정기간 동안 정기적으로 실행하는 통계조사에서는 조사업무를 원활하게 수행할 수 있는 조사원의 수와 배정된 예산에 의해서 표본조사구의 수가 결정될 수밖에 없다.

그러나 조사구 내에서 몇 가구를 표본가구로 선정할 것인가에 따라서 조사구 수는 조정될 수 있으나 한 조사구 내에서 관찰되어야 할 최소한의 실업자 수를 고려한다면, 한 조사구에서 24가구 정도는 조사되어야 추정 및 통계분석에 무리가 없을 것이다.

특히 조사구별로 가중값이 부여되기 때문에 조사구 내에서 선정하는 표본가구수는 일정규모 이상이 되어야 할 것이다. 조사방법을 미국이나 캐나다와 같이 방문조사와 전화 또는 우편조사(인터넷 조사)를 병합하여 적용한다면 조사원 당 전담할 수 있는 조사구의 수를 늘릴 수 있을 것이다.

이 경우도 역시 조사환경과 조사원의 경험 등을 고려하고 또한 방문 면접조사를 다른 조사방법으로 대체할 경우에 생길 수 있는 비표본오차에 대

한 보완방안도 연구되어야 한다.

만일 일부 시군구일지라도 시군구의 실업통계의 작성을 전제로 한다면 표본조사구 수와 표본가구 수는 현재보다는 증가되어야 할 것이나 얼마나 증가시켜야 할 것인가에 대해서는 분석단위에서 언급한 것과 같이 어떤 규모의 시군구까지 실업통계를 생산할 것인지를 우선적으로 결정해야 한다.

전체적인 표본조사구의 수가 결정되면 각 층별로 표본조사구를 배분하는 방법을 두 가지 관점에서 생각할 수 있다.

첫째는 하향식 표본 배분방식으로 일반적으로 통계조사론의 교재 등에서 언급되는 추정량의 분산을 최소화 할 수 있도록 층의 크기와 분산이 큰 층에서는 표본조사구를 많이 배분하고 층의 크기와 분산이 적은 층에서는 적게 배분하는 최적배분법이다. 전국 단위 또는 시도 단위의 실업통계의 추정값에 대한 추정오차만을 고려할 때 적용될 수 있는 표본 배분방식이지만 실제 시군구 실업통계를 생산하고자 할 때 하향식의 최적배분법을 적용한다면 시군구별로 표본조사구의 배분이 불균형적으로 될 가능성이 크다.

두 번째는 상향식 배분방식으로 시군구의 실업통계를 생산하는데 일정 수준의 신뢰도를 유지하도록 하는 최소로 필요한 표본조사구를 배정하는 형식이다. 이때는 전체적인 표본크기를 통제할 수 없을 뿐 아니라 광역시와 도별 실업통계의 신뢰수준이 달라지기 때문에 시도 간 실업통계의 비교분석에 어려움이 있을 수 있다.

실제 표본조사구를 배분할 때 적용할 수 있는 방안으로 하향식 최적배분법과 상향식 배분법을 절충하는 방법을 사용함이 타당할 것이다. 유사한 절충방식을 캐나다의 노동력조사 표본설계에서 적용하고 있다. 먼저 70% 이상의 표본조사구를 하향식 최적배분 방식으로 각 층별로 표본조사구를 배정하

여 표본조사구를 선정한 후에 각 시군구별로 추출된 조사구의 분포를 분석하여 시군구 단위의 실업통계를 생산하는데 추정값의 신뢰도에 문제가 있을 정도로 적게 표본조사구가 배정된 시군구에는 아직 배분되지 않은 조사구를 추가로 배분하는 방식이다.

매월 실업통계를 생산하여 발표한 시군구의 표본조사구의 최소한 크기는 모집단의 구조형태와 층화방식과 밀접한 관계가 있지만 우선 정책적으로 추정값의 목표정도를 정해야 할 것이다.

캐나다의 경우에는 소지역 추정법에 의해서 추계되는 추정값의 허용오차를 최대로 25% 상대표준오차까지 받아들이고 있으나 이는 통계 전문가와 고용통계 전문가들이 협의를 통해서 결정될 문제이다.

시군구의 인구수와 산업적인 중요도가 낮은 경우에는 분기별로 공식적인 실업통계를 생산 발표하여 통계의 신뢰성과 경제성을 높이는 방법도 함께 연구되어야 할 것이다.

표본조사구의 절충식 배분에서 표본조사구의 증감이 추정값의 정확도에 미치는 영향을 살펴보면 인구 규모가 큰 시군구에는 많은 수의 표본조사구가 배정될 것이며 조사구가 많은 시군구에서 몇 개의 조사구를 줄이거나 증가시킨다고 해서 추정값의 정확도에 큰 영향을 미치지 않을 것이지만 인구 규모가 작아서 표본조사구가 적게 배분된 시군구에서는 적은 수의 조사구를 추가할 경우에 시군구 실업통계의 추정값의 정확도는 현저하게 증가할 것이다.

6.2.3 표본관리와 조사방법

현재 경제활동인구조사는 고정표본관리제도를 적용하고 있으나 모집단의 변동을 적시에 반영하지 못하고 표본조사 대상 가구의 응답부담이 크다는 문제점들이 지적되었으나 조사원들이 조사업무를 수행하고 표본가구를 관리하는데 용이하다는 장점도 있을 수 있다.

그러나 정확하고 체감성이 큰 실업통계를 작성하기 위해서는 앞에서 언급된 이점들은 설득력을 잃게되고 고정표본관리제도를 개선해야 한다는 의견을 개진하는 사람들도 적지 않다. 연동교체표본관리제도를 적용하거나 고정표본관리제도를 적용하든 조사원들이 조사대상 가구를 관리하는 방법에서 발생하는 비표본오차는 큰 차이가 없을 것이다. 그러나 조사방법과 조사원의 업무부담에는 큰 차이가 있다.

표본조사구의 관리를 고정표본제도로 하든 연동교체표본제도로 하든 해당 월의 조사구와 바로 전월의 조사구에는 중복되는 부분이 최소한 60%이상이 될 것이며 이와 같이 반복 조사한 표본 조사구에 대해서 해당월의 추정량을 산정할 때 바로 전월의 자료를 이용하는 추정법을 활용하여 추정값의 정확도를 높여야 할 것이다. 이 중복된 표본 조사구의 추정법은 다음 절에 상세하게 다룰 것이다.

경제활동인구조사의 표본 조사구를 고정표본관리제도로 현재와 같은 방법을 유지할 것인가 또는 미국이나 캐나다와 같이 연동교체표본관리제도를 적용할 것인가에 대해서는 여기서 언급은 하지 않겠지만 실업통계의 추정법을 결정하는데는 핵심적인 연구사항이 될 것이다.

미국, 캐나다와 일본 등에서의 노동력 표본조사에서 표본가구를 조사하는 방법은 표본조사대상가구에 대한 기본사항을 파악하고 조사내용에 익숙해지면 방문면접조사를 하지 않고 전화조사 또는 우편조사로 대체하는 형식이다.

그러나 우리나라에서는 매월 조사원이 표본구를 방문하여 면접조사를 하고 있다. 물론 우리나라의 응답률은 99% 이상으로 거의 완벽한 표본조사를 수행하고 있으나 미국과 캐나다의 노동력 표본조사의 응답률이 80% 미만이기 때문에 무응답에 대한 대체추정기법이 발달하여 추정값의 정확도를 유지하고 있다. 그러나 이제부터 직업이 다양화되고 여성들의 사회진출이 확대되면서 조사환경이 어려워지고 있으므로 조사방법에 대한 심층적인 연구가 있어야 할 것이다.

개인적인 제안을 하자면 분기에 1회 정도는 표본가구를 방문하여 면접조사를 현재와 같은 방법으로 수행하고 나머지 달에는 전화조사 또는 인터넷 조사를 도입함이 효과적이고 경제적일 것이라고 생각된다. 조사방법의 개선에 대한 이론적이고 실험적인 연구를 통해서 방법을 개발하고 개선된 방법을 획일적으로 적용하기보다는 도시지역과 농촌지역을 구분하여 다양한 조사방법을 점진적으로 적용하는 방법이 바람직할 것이다.

6.3 추정법

현재의 경제활동인구조사 체계에서 추정 방법 중 보완 개선해야 할 사항을 포함하여 시군구 실업통계 작성에서 보완되어야 할 필요가 있는 내용을 중심으로 설명하겠다.

6.3.1 가중값의 세분화

전국단위 또는 도단위의 실업통계 작성시 현재는 조사구에 가중값을 부여하여 추계하는 방법을 사용하고 있으며 가중값을 동부와 읍면부를 구분하

여 25개 지역층에서 성별을 구분함으로써 50개의 가중값을 산정하였으나 연령계층별과 학력별로 세분화된 경제활동 인구를 추정할 때에는 오차가 발생할 수 있으므로 이와 같은 비표본 오차 발생을 줄이기 위해서는 좀더 세분화된 가중값 부여 방안을 연구해야 할 것이다.

세분화 가중값을 산정하기 위해서는 모집단의 구조를 세분화하여 분석해야 하고 또한 상주인구 추계시에도 세분화된 범주에 따라서 추계인구를 계산해야 할 것이다.

2000년 센서스 자료를 분석하여 성별-연령계층별-학력별로 세분화하여 범주를 구분하고 각 범주별로 모집단의 구조적 특성을 파악하며 또한 상주인구 추계시에도 세분화된 범주별로 추계인구를 계산해서 매월 조사구별 가중값을 산출하여 이용해야 할 것이다. 물론 현재의 25개 지역층을 성별-연령대별-학력별로 세분화 한다면 750개로 세분화된 가중값을 산정하여 추정에 반영함으로써 좀더 안정적인 실업통계를 생산할 수 있을 것이다.

6.3.2 고정표본관리제도의 추정법 개선

동일한 가구를 매월 조사한다면 각 조사구별로 해당월과 전월간의 관찰값(실업자, 취업자 등)들은 높은 상관관계를 가질 것이다. 해당월의 실업통계를 추정할 때 전월들의 자료 또는 추정값을 이용한다면 추정값의 정확도는 높아질 것이다.

예를 들어 조사구 단위로 본다면 일종의 시계열 자료 형식이 될 것이다. 적합한 시계열 모형을 구축하여 해당월 실업통계의 예측값(predicted value)을 계산할 후에 해당월의 조사된 자료에서 실업통계의 추정값을 계산한다면

두 개의 추정값들을 얻을 수 있다. 이들을 선형결합한다면 해당월의 자료만 이용한 추정값보다 정확도가 높은 추정값이 얻어질 것이다.

이에 관한 내용은 캐나다 통계청에서 A-K 추정법을 개발하면서 연구하였기 때문에 어렵지 않게 우리나라의 통계조사 환경에 맞는 추정량을 개발할 수 있을 것이다.

6.3.3 무응답 대처방안 연구

현재와 같이 거의 완벽한 응답률을 유지할 수 있다면 다행이지만 앞으로는 조사여건이 나빠지면서 단위 무응답 또는 항목 무응답이 발생할 가능성이 높아지고 있다. 특히 현행 방문 면접조사법을 개선하여 방문면접조사와 전화조사, 방문면접조사와 우편(인터넷)조사를 결합한 조사방법을 적용하게 된다면 무응답자는 현재보다는 많이 발생할 것이다. 물론 무응답자를 줄이고자 하는 노력이 우선 되어야 하겠지만 불가항력적으로 무응답자가 발생했을 경우에 대처하는 방안이 연구되어야 할 것이다.

조사단위가 무응답일 경우에는 가중값 조정방법이나 부차표본추출법을 적용하여 비표본오차와 편향을 줄이도록 노력하고 있으나 항목 무응답의 경우에는 조사단위의 특성을 파악하여 유사한 조사단위의 응답내용을 대입하는 대체(Imputation) 방법이 사용되기도 한다.

항목무응답에 대한 대체방법은 평균대체, 최근방대체, 콜드덱(Cold Deck), 회귀대체, 이월대체(Carry-Over Imputation), 핫덱(Hot Deck), 베이지안 대체와 복합대체 등이 있으나 몇 가지 대표적인 대체 방법만 간단하게 설명해보자.

(1) 평균대체(Mean Imputation)

전체 표본을 몇 개의 대체층으로 분류한 다음에 각 층에서 항목 무응답이 있을시에는 응답한 내용들의 평균을 계산하여 대입하는 형식이다. 해당 층에서 항목 무응답이 많거나 응답자와 무응답자 간의 차이가 있을 경우에 추정에서 편향이 발생할 수 있으나 적용방법이 쉽고 평균이나 합계의 모수추정에서는 점추정량의 편향을 감소시키는 효과가 크다.

(2) 최근방대체(Nearest Neighbor Imputation)

미국의 경상인구조사(CPS : Current Population Survey)에서 결측치에 대한 대체방법으로 사용되고 있는 형식이며, 전체 표본을 대체층으로 구분한 후에 각 층 내에서 응답자료를 순서대로 정렬하여 결측값이 있는 항목은 바로 그 이전 항목의 값을 삽입하는 형식이다. 응답된 집단에서 무응답한 자료의 특성과 가장 유사한 자료값을 찾아서 대체하므로 편리성이나 비용적인 측면에서 유용하게 이용되고 있다.

(3) 회귀대체(Regression Imputation)

무응답이 있는 항목 y에 응답이 주어진 항목들($x_1, \cdots, x_k$)을 y의 보조변수로 생각하고 회귀모형을 적합시켜서 적합된 모형을 이용하여 추정한 값을 대체값으로 대입하는 형식이며 i번째 결측값에 대한 회귀모형은 아래와 같이 고려할 수 있다.

$$\hat{y_i} = \hat{\beta_0} + \sum_{k=1}^{k} x_{ki}\hat{\beta_k} + e_i$$

여기서 β_i 는 응답한 자료를 이용하여 최소제곱 추정법으로 계산한 회귀계수 벡터이다. 이 회귀대체법은 미국의 경상조사자료에서 적용하여 다른 대체방법과 비교했을 때 평균절대편차(Mean Absolute Deviation)을 가장 적게 갖는 것으로 입증되었다.

(4) 핫덱대체(Hot Deck Imputation)

조사된 자료를 대체층으로 구분하고 대체층 내에서 대체값을 확률추출법으로 랜덤하게 선택하여 항목무응답에 대체하는 형식이며 평균대체방법이 y의 분포를 왜곡시킬 수 있다는 문제점을 완화할 수 있는 잇점이 있다.

이상에서 4종류의 대체방법에 대해 간단하게 설명하였다. 앞에서 언급된 무응답 대체방법 중에서 어느 것을 선택할 것인가를 결정하기 전에 무응답에 대한 본질적인 연구가 있어야한다.

모집단에서 응답자와 무응답자간의 차이가 있는지, 있으면 얼마나 되며 어떻게 측정해야할 것인가를 연구하는 것을 동일성연구라고 한다. 동일성연구에 대한 일반적인 내용을 소개하는 것도 무응답의 대체방안에서 기초연구가 될 것이므로 여기에서 살펴보자.

(5) 동일성 분석

동일성분석에서는 조사대상의 응답확률이 주 연구변수와 어느 정도 관련이 있는지 또는 응답자들이 무응답자들과 어떻게 그리고 얼마나 다른지를 측정함으로써 무응답 편향의 가능성을 밝히는 것이다. 이러한 동일성 분석연구를 기반으로 해서 통계 분석가들은 조사모집단들 사이의 응답패턴을 연구

하여 추정값에 대한 무응답의 영향을 양적으로 평가할 수 있다. 마찬가지로, 이들 연구들은 응답자와 무응답자들을 비교하거나 응답자들을 전체 모집단과 비교함으로써 수행될 수 있다. 응답 표본집단과 무응답 표본집단의 분석에서는 사회인구학적특성(예를 들면, 연령, 인종, 성별, 그리고 교육)들이 일반적으로 유용하게 사용된다.

표본 부집단들 간의 응답률 차이를 연구하는데 있어서, 응답확률과 연구변수(Y)수와 상관이 높은 보조변수와의 관계를 결정하는 것이 연구의 핵심이다. 부차집단의 응답률이 통계적으로 보조변수의 값들과 연관이 있다는 것은 Y변수를 사용해서 얻은 추정값들이 무응답을 원인으로 하여 편향된다는 것을 가리킨다. 이러한 편향이 생기는 이유는 응답률과 보조변수 값과의 연관성이 개개의 응답확률과 Y변수 값간의 상관을 0이 되지 않게 하는 경향이 있어 결국에는 추정값에 편향이 생기기 때문이다. 더 직관적으로 말하면, 응답률과 보조변수간의 관계는 표본 응답자들간의 Y변수와 보조변수의 분포들이 전체 표본에서의 대응분포와 같지 않게 된다는 것을 의미한다. 응답자들간에 Y변수의 분포가 무응답으로 인해 치우치면, 응답표본으로부터 생산된 추정값들도 마찬가지로 어느 쪽으로인가 편향이 생기게 될 것이다.

동일성 분석의 두 번째 형태는 보조자료를 이용해서 응답자들과 무응답자들을 비교한다. 보조변수에 대한 간단한 기술적 척도들이(예를 들면, 평균) 일반적으로 이들의 비교를 위해서 이용된다. 보조변수를 선택하기 위한 기준은 부집단의 응답률에 대한 연구에서와 같다. 지역표본에서, 지리적 영역, 인구밀도, 관측된 인종, 추출된 지역단위의 분포들이 응답자들과 무응답자들을 위해서 비교될 것이다. 다른 표본에서, 기술적 척도나 사회인구학적 변수들의 분포를 대비하여 분석한다.

동일성 분석의 세 번째 형태는 무응답자들이 얼마나 닮았느냐의 정도에

따라 다른 응답자들의 부그룹을 비교함으로써 주요 연구변수들에 관한 응답자들과 무응답자들 간의 차의 징후를 찾는 것이다.

무응답 문제를 다루는 이러한 방법은 때때로 자료수집이 여러 주기(waves)에 걸쳐 완성되는 우편조사와 인터넷 조사등 에서 사용된다. 첫 번째 주기는 첫 번째 우편 발송과 첫 번째 추가 권유 사이의 기간이고 두 번째 주기는 첫 번째 추가 권유와 두 번째 추가 권유 사이의 구간이다. 이런 식으로 주기를 정한다. 주기를 사용하는 근본적인 이유는 c번째 주기에서 응답하는 표본구성원은 c가 증가할수록 참가하는 것을 더 꺼린다는 것을 의미한다. 면접조사에서 주기에 대해 비교 가능한 척도는 참여를 얻는데 필요한 방문의 횟수가 될 것이다.

응답자 자료와 주기 또는 응답을 하게 한 방문은 단위무응답의 편향효과를 다루고 확인하기 위해 여러 분야에서 연구되고 사용되어 왔다. 이러한 방법을 고려해서 발표된 대부분의 것들은 우편조사에도 적용되어 왔다. 어떤 경우에는 각 주기로부터 개별적으로 얻어진 추정값들로부터 관측된 경향을 바탕으로 나중에 응답한 사람들이 처음에 응답한 사람들과 다른 경향이 있는 지를 조사한다. 각 주기를 통해 합쳐진 자료를 사용한 추정값은 같은 이유에 대해 비교할 수 있다. 응답 주기가 응답확률에 대한 적절한 표현이라고 가정하면, 주기에 의한 추정값들 간의 단조 증가나 단조 감소하는 경향은 개개의 응답확률과 Y변수와의 상관을 나타낼 것이다. 그래서 무응답으로 인해 추정값이 편향되는 경향이 생기게 된다.

자료수집기간 동안에 얻어진 자료에 따라 응답자를 분류함으로써 얻어진 일련의 추정값들은 완전 응답이 얻어졌을 때의 추정값을 예측하는데 사용된다. 외삽법은 위에서 언급했던 일련의 추정값들을 가정한 모형에 적합시키는 보정과정을 의미한다.

그리고 적합된 모형은 외삽법에서 기본모형으로 사용된다. 이들 모형에서 독립변수(x)는 설문지가 완성됐을 때 자료를 수집하는 동안의 점수(예를 들면, 방문 횟수, 설문 작성 시간, 누적 응답률)를 측정하고, 종속변수(y)는 x에 의해 결정된 점수를 통해 수집된 자료들을 모아서 얻은 조사추정값(예를 들면, 평균)이다.

대부분의 외삽법의 기본개념은 같다. 응답자들이 무응답자들과 얼마나 닮았는가를 나타내주는 응답결과변수(보조변수 X로 간주함)라 불리는 것이 각 응답자들에게 주어진다. 응답결과변수를 하나의 주요 연구변수(Y변수)에 연결시켜주는 가정된 모형은 응답자료를 이용해서 만들어진다. 마지막으로, 적합된 모형은 최종결과를 보정하는데 사용된다. 보정방법은 X변수에 사용된 척도에 따라 다르고, 통계적 모형은 X변수와 Y변수의 관계를 연관시켜 줄 수있다고 가정한다.

우편조사에 적용되었던 외삽법의 개념을 사용한 최초의 논문중의 하나에서, Hendricks(1949)는 다음과 같은 로그-선형모형을 적합시켰다.

$$y = \alpha x^{\beta}, \tag{8.1}$$

여기서 x는 응답을 얻는 데 필요한 방문 횟수, 그리고 α와 β는 상수로서 응답자들의 자료로부터 추정되는 회귀모수이다. Hendricks가 사용한 x의 정의는 Scott(1961)가 요약한 초기의 전형적인 외삽법이다. Filion(1976)과 Jones(1983)는 다음의 선형모형을 가정하였다.

$$y = \alpha + \beta x,$$

여기서 x는 주어진 주기까지의 누적응답률이다. Von Riesen과 Novotny(1979)는 x에 대해서 조사 완료시간을 사용하였고, 자료를 수집하

는 동안 모형의 선형성에 변화가 있을 것이라는 가정을 반영하기 위해서 구분적 선형모형(piecewise linear model)을 적합시켰다. 하나의 기울기 변화를 갖는 우편조사에서는 다음과 같은 모형을 사용한다.

$$y = \alpha + \beta_1 x + \beta_2 (x - X_t)\delta,$$

이때, X_t는 기울기가 β_1에서 $\beta_1 + \beta_2$로 변하고 δ는 다음과 같은 x척도상에 있는 점이다.

$$\delta = \begin{cases} 1, & \text{만약 } (x - X_t) \geq 0 \\ 0, & \text{그밖의 경우.} \end{cases}$$

마지막으로, Ognibene(1971)은 y에 x(누적응답률)를 관련시키는 여러 모형들의 편리성을 비교하는 연구에서 다음과 같은 쌍곡선모형이 가장 유용한 결과를 제공한다고 결론지었다.

$$y = \alpha + \frac{\beta}{x},$$

x에 대한 정의가 외삽을 위해서 사용된 적합된 모형을 결정한다. x가 누적응답률일 때, 외삽에 의한 추정값은 완전응답의 경우(누적응답확률 $x = 1$) Y변수에 대한 것이 된다. 그런데, x가 응답을 얻기 위해 투입한 노력(주기, 방문 횟수, 또는 설문을 완료하는 데 걸린 시간)의 정도에 대한 척도일 때, 외삽의 목적을 명확하게 해두어야 한다. 무응답에 대한 Y변수의 값을 적절하게 예측할 응답결과변수를 임의로 선택해야 한다.

조사무응답을 다루는데 외삽의 유용성이 명확하게 입증되지 않았다. 외삽을 옹호하는 사람들은 이 방법이 비교적 단순하고 조사무응답을 다루는데 비용이 적게 드는 방법이라고 주장한다. 외삽의 결과들은 무응답 부차표본을 뽑는 비용이 소요되지 않거나 복잡한 재가중치방법을 사용하지 않고도 얻어

질 수 있다. 반면에, 존재하는 증거들에 비추어볼 때 외삽을 싫어하는 사람들은 설문을 완료하는데 걸린 시간의 측정과 중요연구변수간의 연관 정도가 명확하지 않다고 말한다.

실제적인 이점에도 불구하고, 외삽법은 Y변수의 값을 보다 잘 예측하는 x의 척도가 발견될 때까지는 널리 사용되기 어려울 것으로 보인다. 많은 척도들이 갖는 문제는 그들이 응답자들의 참여확률을 나타내지 못한다는 것이다. 예를 들면, 면접조사에서 응답을 얻는데 요구되는 노력의 양이 단지 응답자의 외형적인 태도를 반영하는 것이지 참여하려는 의지나 능력을 반영하는 것은 아니기 때문이다. 추측 상, 하나의 이상적인 응답척도는 이들 세 개의 특성을 모두 반영해야 한다. 우편조사에서, 응답을 하는 주기와 설문을 완료하는 시간은 질질 끌거나 잊어버리는 응답자들의 성향을 가리키는 것으로 생각할 수 있기 때문이다.

동일성분석은 지속적이고 심층적으로 연구되어야하고 특히 대규모통계조사에서는 필수적으로 해결되어야할 과제이다. 본 연구에서는 적절한 대체방법을 선택하여 적용하는 것을 중심으로 언급하였으나 실제 통계생산에서는 대체한 후에 통계분석과 추정량의 분산은 어떻게 계산해야 하는지에 대해서 좀더 심층적인 연구가 필요할 것이다.

6.3.4 소지역 추정법 연구

시군구의 실업통계 작성을 목적으로 연구한 소지역 추정법은 다른 소지역인 교차분류된 성별-산업별 범주에 대한 소지역 추정법으로 적절하지 않을 경우가 많다. 따라서 소지역 추정법은 연구주제 또는 통계생산단위에 따라서 다르게 적용될 수 있으므로 다양한 추정법에 대해서 폭넓게 연구되어

야 할 것이다.

본 연구에서는 직접 추정법, 합성 추정법과 복합 추정법에 대해서만 수치적으로 비교분석하여 복합 추정법이 가장 안정적이고 정확성이 크다는 것을 보였으나 전 월들의 조사자료를 함께 이용하거나 중앙고용정보원의 구직등록 데이타베이스와 실직보험 청구정보를 이용할 경우에는 또 다른 형태의 소지역 추정법이 연구되어야 할 것이다.

우선 먼저 2003년의 경제활동 인구조사의 표본설계에 대한 총괄적인 연구방향이 설정된 다음에는 주어진 여건에 적합한 소지역 추정법의 연구와 컴퓨터 프로그램의 개발이 진행되어야 할 것이다.

통계 선진국인 미국에서 사용하는 소지역 추정 모형은 시계열 및 횡단면 모형을 결합한 형식이지만 우리나라의 경우에는 아직까지는 경제적 상황이나 노동시장의 여건이 안정적인 변동상태를 보이지 못하므로 시계열-횡단면 모형은 순수 연구 목적에서 참고 대상으로 삼을 수 있으나 실용화하기 위해서는 장기간의 이론적인 내용과 실제로 이용할 보조정보의 준비가 선행되어야할 것으로 생각된다.

그러나 캐나다에 적용하고 있는 복합 추정법은 우리나라에서 Bench Marking 모형으로 삼고 연구한다면 우리나라의 고용관련통계에 적용할 수 있을 것으로 사료된다. 2000년 인구주택총조사의 자료 중에서 10%표본조사 자료를 대상으로 연구되어야할 내용을 요약하면 다음과 같다.

(1) 모집단 층화

복합 추정법을 시군구의 실업통계 생산에 적용하기 위해서 편향이 적은 합성 추정량을 계산해야하는데 합성 추정법에서 가정한 전제조건을 충족

시켜줄 수 있도록 전국을 광역층으로 분할하고 광역층 내에서는 세분화된 범주의 실업률을 또는 경제활동참여율 등의 노동력에 관련된 특성이 거의 동일하도록 모집단을 층화하여야 한다.

예를 든다면 전국을 7대 광역시와 경기도, 가원도, 충청도, 전라도와 경상도로 12개 지리적인 광역 권역으로 구분하고 5개 도단위 권역은 시지역과 읍면지역으로 분할하여 전국을 18개 지역층으로나눈후에 각 지역층에서 실업률이 가장 유사한 성별-연령대별 범주(남-여, 34세이하-35세이상)로 구분하는 절차를 카이제곱 검정을 이용하여 수행할 수 있는 프로그램을 개발하여 시뮬레이션 연구를 통해서 최적 층화를 실행한다.

위와 같은 전산프로그램을 이용한 층화작업은 2만여 개의 표본조사구를 대상으로 진행되어야하므로 컴퓨터 전문지식과 실무경험을 겸비한 연구진과 소지역 통계전문가들의 협동연구로 진행되는 것이 바람직하다.

만일에 18개 지역층에서 4개 범주로 구분하는 층화 방안을 적용한다면 각 범주의 셀에는 평균적으로 20개 정도의 조사구가 배정되므로 직접 추정법에 의해서 실업률을 산정하더라도 안정적이고 상당한 수준의 신뢰성을 갖출 것으로 생각된다.

(2) 표본크기 결정 및 표본배분

표본가구의 크기는 조사인력과 배정된 예산에 의해서 결정되기 때문에 표본가구수를 33,000가구 정도로 가정해도 무리는 아닐 것이다. 현재와 같이 한 표본조사구에서 8가구를 한구역으로한 인근 3개 구역을 조사대상 가구로 할 경우에는 1,340개 정도의 표본조사구를 1차 추출단위에 대한 표본크기로 생각할 수 있다. 만일에 행정자치부 또는 노동부와 같은 부처에서 필요한 소

지역의 고용통계를 원만하게 공급받는다는 조건으로 해당부처에서 재정적인 분담금을 지원할 수 있다면 캐나다와 유사한 방법으로 통계청의 표본조사구의 크기에 추가하여 타부처의 분담금에 해당하는 표본조사구를 보충하여 소지역 통계의 신뢰성을 높이는 방법을 생각할 수 있을 것이다.

전국적으로 표본조사구가 결정되면 각층별로 표본을 배분하는 방안이 연구되어야하지만 18개 지역층에 대해서 표본을 먼저 배분할 것인가 아니면 행정구역별(16개 시도단위)로 표본을 배정할 것인가를 결정하는 방법이 연구되어야한다. 각 층별 표본배분은 실업통계 생산단위를 먼저 검토한 후에 결정되어야할 것이다. 만일에 7대 광역시에서는 구별로 실업통계를 매월 발표하고 도지역에서는 인구 10만명 이상의 시군에 대해서는 매월 실업통계를 생산해서 발표하며, 나머지 시군에 대해서는 분기별로 실업통계를 작성하여 공표 한다면 표본배분은 상향식 배분법과 최적 배분법을 결합하는 형식의 배분법을 적용할 수 있을 것이다.

먼저 인구 10만명 이상의 시군과 광역시의 구에 10개(가정한 숫자임) 표본조사구를 배분한 후에 나머지 표본조사구를 18개 지역층에 대해서 최적 배분법 또는 지역층의 크기(가구수)에 비례하도록 비례배분법을 적용하여 일차로 시군구까지 표본조사구를 배분한 후에 시군구별로 배정된 표본조사구의 분포를 분석하는 반복적인 과정을 통해서 표본배분을 수행한다면 적절하게 표본조사구의 배분이 이루어질 것이다.

(3) 가중값 산정과 추정

현재와 같이 조사구별로 가중값을 산정하는 방법만 적용할 것이 아니라 성별-연령대별로 가중값을 부여하는 방법도 이론적인 측면에서 뿐만아니

라 컴퓨터 프로그램을 통해서 계산의 용이성을 고려한 가중값 부여 방안도 연구되어야 다양한 교차분류에 대해서도 신뢰성을 갖춘 실업통계의 생산이 가능해지고 통계의 실무 활용성이 제고될 수 있을 것이다.

이와 같은 다양한 가중값 산정방법의 연구는 계산의 편이성도 고려되어야 하겠지만 보다 더 주의해야할 사항은 교차분류의 범주를 다양하게 했을 때도 추정값의 편향이 최소화되어야 할 것이다.

현재 통계청이 경제활동인구조사에서 사용하는 가중값 산정은 무응답을 전혀 고려하지 않았지만 앞으로는 무응답이 발생할 경우에 대비해한 가중값 산정방안이 연구되어서 무응답 발생이 고려된 가중값의 산정 알고리즘이 개발되어야 할 것이다.

추정에서는 표본조사구의 수가 충분히 큰 경우에는 직접 추정법을 적용하여 실업통계를 생산할 수 있으나 표본조사구의 수가 적을 경우에는 조사모집단의 구조분석을 통해서 별도의 층화작업을 한 후에 복합 추정법을 이용하여 시군구의 실업통계를 생산해야 할 것이다.

복합 추정법에서 핵심적인 사항은 직접 추정량과 합성 추정량을 결합하는 가중값을 결정하는 것이다. 가중값을 결정하는데 각 소지역별로 특성을 고려한 소지역별로 상이한 가중값을 계산하는 방법과 소지역별 특성보다는 권역별 평균제곱오차를 최소로 하는 가중값을 산정하는 방법이 있지만 두가지 방안 모두가 합성 추정량의 평균제곱오차의 추정값을 필요로 하기 때문에 근사적인 가중값을 산정하는 방법을 연구하여 적용하고 있다.

복합 추정법으로 산정한 시군구의 실업통계는 각 시군구의 특성만을 고려했기 때문에 각 행정구역별(광역시와 도단위)로 시군구의 실업자 추정값을 합산했을 때 각 행정구역별로 직접 추정법으로 추계한 추정값과 일치해

야할 것이며 만일에 일치하지 않는다면 각 행정구역별로 직접 추정한 추정값과 해당 시군구의 추정값을 합산한 간접 추정값이 같도록 조정해야 할 것이다.

완전한 추정법이 개발된 후에는 실제로 조사모집단(10% 표본조사구의 집합)에서 표본을 추출하여 실업통계를 생산하여 모수의 참값과 추정값간의 오차분석을 통해서 표본설계와 추정법의 타당성을 검증해야 할 것이다.

우리나라의 연구진에 의해서 독자적인 소지역 추정법의 연구보다는 캐나다의 통계청을 방문하여 캐나다 소지역의 노동력 통계를 작성하는 방법을 전수 받을 수 있다면 효과적이고 실용적인 연구를 위해서 바람직할 것이며 시군구 실업통계를 생산하는 기법이 개발되고 체계가 정립된다면 우리나라의 다양한 분야에서 통계수준이 한 차원 더 향상될 것으로 생각된다.

소지역 추정법 중에서 전제 조건만 충족된다면 합성 추정법이 개념적으로나 방법론적으로 가장 유용한 추정법이지만 전제조건이 맞지 않을 경우에는 편향 추정량이 되므로 평균제곱오차(MSE : Mean Squared Error)를 계산해야 하는데 평균제곱오차를 정확하게 계산할 수 있는 방법의 연구는 소지역 추정법의 큰 연구과제이다. 소지역 추정법에서 평균제곱오차를 추정하는 방법에 대해서 지금까지 연구된 내용을 요약하고 연구방향을 제시하고자 한다.

6.4 소지역 추정법에서 평균제곱오차의 추정

앞에서 복합추정법을 적용한 시군구 실업통계의 작성 가능성을 이론적인 측면에서뿐만 아니라 수치적인 시뮬레이션을 통해서 검증을 하였으나 복합

추정량의 평균제곱오차의 추정에 대해서는 상세하게 언급하지 못하였다. 본 절에서 소지역 추정법에서 평균제곱오차의 추정에 대해서 살펴보고자 한다.

복합추정량의 형태는 직접추정량과 합성추정량의 선형결합으로만 주어지는 것이 아니고 합성추정량의 형식에 따라서 다양하게 주어질 수 있다. 일반적으로 복합추정량은 다음과 같은 형식으로 표현되는 경우가 많다.

$$\hat{Y}_i^C = \omega_i \hat{Y}_i^D + (1-\omega_i) \hat{Y}_i^S , \qquad (6.1)$$

여기에서 $\hat{Y}_i^D$는 i번째 소지역의 직접추정량이고, $\hat{Y}_i^S$는 일종의 합성추정량으로 다양한 형태로 주어질 수 있다.

식 (6.1)에서 가중값 ω_i를 결정하는 방법은 $\hat{Y}_i^C$의 평균제곱오차를 최소로하는 방안과 모든 소지역들에 대해서 평균제곱오차의 평균을 최소로하는 방안을 고려할 수 있다.

먼저 $MSE(\hat{Y}_i^C)$를 최소로하는 ω_i는 다음과 같은 방법으로 계산할 수 있다.

$$\begin{aligned} MSE(\hat{Y}_i^C) &= E\left[\omega_i \hat{Y}_i^D + (1-\omega_i) \hat{Y}_i^S - Y_i\right]^2 \\ &= E\left[\omega_i(\hat{Y}_i^D - Y_i) + (1-\omega_i)(\hat{Y}_i^S - Y_i)\right]^2 \\ &= \omega_i^2 Var(\hat{Y}_i^D) + (1-\omega_i)^2 MSE(\hat{Y}_i^S) \end{aligned} \qquad (6.2)$$

식 (6.2)는 ω_i에 대해서 2차식의 형태이므로 ω_i에 대한 최소값은 (6.2)식을 ω_i로 미분하여 아래 방정식의 해를 구하면 된다.

$$\frac{\partial MSE(\hat{Y}_i^C)}{\partial \omega_i} = 2\omega_i Var(\hat{Y}_i^D) - 2(1-\omega_i) MSE(\hat{Y}_i^S) \qquad (6.3)$$

식 (6.3)에서 해를 구하면 ω_i의 최적값은 다음과 같다.

$$\omega_{i(opt)} = \frac{MSE(\hat{Y}_i^S)}{MSE(\hat{Y}_i^S) + Var(\hat{Y}_i^D)} \tag{6.4}$$

최적 가중값 $\omega_{i(opt)}$은 모수($MSE(\hat{Y}_i^S)$와 $Var(\hat{Y}_i^D)$)를 포함하고 있으므로 다음과 같은 식으로 추정할 수 있으나 소지역에 따라 매우 불안정한 특성을 갖고 있는 문제가 있다.

$$\begin{aligned}\hat{\omega}_{i(opt)} &= \frac{mse(\hat{y}_i^s)}{(\hat{y}_i^s - \hat{y}_i^D)^2} \\ &= \frac{(\hat{y}_i^s - \hat{y}_i^D)^2 - v(\hat{y}_i^D)}{(\hat{y}_i^s - \hat{y}_i^D)^2} \\ &= 1 - \frac{v(\hat{y}_i^D)}{(\hat{y}_i^s - \hat{y}_i^D)^2}\end{aligned} \tag{6.5}$$

식 (6.5)의 값은 때로는 음수가 되거나 1보다 큰 수가 나올 수 있으므로 적용하는데 문제가 될 수 있다.

다음에는 모든 소지역들의 평균제곱오차의 추정값들의 평균을 최소로하는 최적 가중값 $\hat{\omega}_{(opt)}$은 다음과 같이 주어진다.

$$\hat{\omega}_{(opt)} = 1 - \frac{\sum_i v(\hat{y}_i^D)}{\sum_i (\hat{y}_i^s - \hat{y}_i^D)^2} \tag{6.6}$$

식 (6.6)은 식 (6.5)에 비해 안정적이기는 하지만 각 소지역별 특성을 제대로 반영할 수 없는 단점을 가지고 있다.

식 (6.2)에서 $MSE(\hat{Y}_i^C)$를 추정하기 위해서는 $MSE(\hat{Y}_i^S)$를 추정해야할

것이다. 일반적으로 $MSE(\hat{Y}_i^S)$를 추정하는 두 가지 방법들에 대해 살펴보자.

첫 번째 방법은 $MSE(\hat{Y}_i^S)$에 대한 다음 (6.7)식을 이용하는 방법이다.

$$\begin{aligned} MSE(\hat{Y}_i^S) &= E(\hat{Y}_i^S - Y_i)^2 \\ &= E(\hat{Y}_i^S - \hat{Y}_i^D)^2 + E(\hat{Y}_i^D - Y_i)^2 + 2E(\hat{Y}_i^S - \hat{Y}_i^D)(\hat{Y}_i^D - Y_i) \\ &= E(\hat{Y}_i^S - \hat{Y}_i^D)^2 - Var(\hat{Y}_i^D) \end{aligned} \quad (6.7)$$

식 (6.7)에서 $MSE(\hat{Y}_i^S)$의 추정값은 각 항목의 추정값들로 나타낼 수 있으므로 다음과 같이 계산할 수 있다.

$$MSE(\hat{y}_i^s) = (\hat{y}_i^s - \hat{y}_i^D)^2 - v(\hat{y}_i^D) \quad (6.8)$$

두 번째 방법은 Marker(1995)가 제안한 방법으로 다음 식을 이용할 수 있다.

$$MSE(\hat{y}_i^s) = v(\hat{y}_i^s) + \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i^s - \hat{y}_i^D)^2 - \frac{1}{m}\sum_{i=1}^{m} v(\hat{y}_i^D) - \frac{1}{m}\sum_{i=1}^{m} v(\hat{y}_i^s) \quad (6.9)$$

여기에서 m은 전체 소지역의 개수이다.

식 (6.9)는 다음과 같이 분산과 편향의 항으로 평균제곱오차를 표현할 수 있다는 사실을 이용한다.

$$MSE(\hat{Y}_i^S) = Var(\hat{Y}_i^S) + [Bias(\hat{Y}_i^S)]^2$$

편향제곱의 평균의 추정량은 다음과 같이 표현할 수 있으므로 추정값은

항목별 추정값을 대입하여 산정할 수 있을 것이다.

$$\frac{1}{m}\sum_{i=1}^{m}[Bias(\ \hat{Y}_i^S)]^2=\frac{1}{m}\sum_{i=1}^{m}MSE(\ \hat{Y}_i^S)-\frac{1}{m}\sum_{i=1}^{m}Var(\ \hat{Y}_i^S)$$

$$=E\Big[\frac{1}{m}\sum_{i=1}^{m}(\ \hat{Y}_i^S-\ \hat{Y}_i^D)^2\Big]-\frac{1}{m}\sum_{i=1}^{m}Var(\ \hat{Y}_i^D)-\frac{1}{m}\sum_{i=1}^{m}Var(\ \hat{Y}_i^S)$$

$$\text{추정값}\Big\{\frac{1}{m}\sum_{i=1}^{m}[Bias(\ \hat{Y}_i^S)]^2\Big\}=\frac{1}{m}\sum_{i=1}^{m}(\ \hat{y}_i^s-\ \hat{y}_i^D)^2$$

$$-\frac{1}{m}\sum_{i=1}^{m}v(\ \hat{y}_i^D)-\frac{1}{m}\sum_{i=1}^{m}v(\ \hat{y}_i^s) \qquad (6.10)$$

위의 (6.10)식을 이용하면 (6.9)식을 얻을 수 있다.

식 (6.1)에서 가중값 ω_i를 산정하는데 설계기반 형식의 방법보다는 모형기반 형식의 방법이 많이 연구되고 있다. 모형기반 형식의 일반적인 모형은 랜덤효과와 고정효과를 병합한 혼합 회귀모형식이 될 것이며 다음과 같은 간단한 형식을 생각해 보자.

$$\hat{Y}_i^D=X_i^T\beta+z_i\ b_i+e_i\ ,\quad i=1,2,\cdots,m \qquad (6.11)$$

여기에서 e_i는 표본오차로서 b_i와는 독립적이고 기대값은 0, 분산은 σ_i^2을 가정하며, b_i는 상호 독립적인 확률변수로서 기대값은 0, 분산은 σ_b^2으로 가정하고 z_i는 알려진 상수로 가정하였다.

식 (6.11)에 주어진 모형의 복합추정량은 다음과 같은 형식이 될 것이다.

$$\hat{Y}_i^{EC}=\omega_i\ \hat{Y}_i^D+(1-\omega_i)X_i^T\ \tilde{\beta} \qquad (6.12)$$

여기에서 $\tilde{\beta}$는 가중회귀계수 추정벡터이며 아래와 같고

$$\tilde{\beta}=\Big(\sum_{i=1}^{m}\frac{X_iX_i^T}{\sigma_b^2z_i^2+\sigma_i^2}\Big)^{-1}\Big(\sum_{i=1}^{m}\frac{X_i\ \hat{Y}_i^D}{\sigma_b^2z_i^2+\sigma_i^2}\Big),$$

또한 가중값 ω_i는 다음과 같다.

$$\omega_i = \frac{\sigma_b^2 z_i^2}{\sigma_b^2 z_i^2 + \sigma_i^2} \ ,$$

여기에서 σ_i^2은 표본오차이므로 안다고 가정할 수 있으나 σ_b^2을 모르므로 이에 대한 추정값을 대입하면 식 (6.12)의 추정량은 경험적 최량 선형불편예측(EBLUP: Empirical Best Linear Unbiased Prediction)추정량이 된다. 앞으로 간편하게 EBLUP 추정량으로 표기하자. σ_b^2을 추정하는 방법은 정규분포를 가정하였을 때 적용할 수 있는 최대우도추정법, Henderson의 상수 적합방법과 적률결합법 등이 있으나 여기서는 언급을 하지 않겠다.

다음에는 EBLUP 추정량의 MSE를 추정하는 방법을 알아보자. EBLUP의 MSE는 σ_b^2과 σ_i^2이 주어졌다는 가정 하에서 발생하는 오차의 제곱과 σ_b^2을 추정하는데 따른 오차 제곱의 항으로 다음과 같이 쓸 수 있다.

$$MSE(\ \hat{Y}_i^{EC}) = g_{1i}(\sigma_b^2) + g_{2i}(\sigma_b^2) + g_{3i}(\sigma_b^2) \ , \qquad (6.13)$$

여기에서 $g_{1i}(\sigma_b^2) = \frac{\sigma_b^2 z_i^2 \sigma_i^2}{\sigma_b^2 z_i^2 + \sigma_i^2} = \omega_i \sigma_i^2$,

$$g_{2i}(\sigma_b^2) = (1-\omega_i)^2 X_i^T \left(\sum_{i=1}^{m} \frac{X_i X_i^T}{\sigma_b^2 z_i^2 + \sigma_i^2} \right)^{-1} X_i \ ,$$

$$g_{3i}(\sigma_b^2) = \frac{\sigma_i^4 z_i^4}{(\sigma_b^2 z_i^2 + \sigma_i^2)^{-3}} \ Var(\ \hat{\sigma}_b^2) \ ,$$

$$Var(\ \hat{\sigma}_b^2) = \frac{2}{m} \sum_{i=1}^{m} \left(\sigma_b^2 + \frac{\sigma_i^2}{z_i^2} \right)^2$$

식 (6.13)에 주어진 $MSE(\ \hat{Y}_i^{EC})$의 추정값은 각 항목별로 추정값을 계산하여 대입하는 방법을 취하면 아래와 같은 결과를 얻을 수 있다.

$$mse(\hat{Y}_i^{EC}) = g_{1i}(\hat{\sigma}_b^2) + g_{2i}(\hat{\sigma}_b^2) + 2g_{3i}(\hat{\sigma}_b^2) \quad (6.14)$$

식 (6.14)의 세 번째 항의 계수가 2인 이유는 $g_{1i}(\hat{\sigma}_b^2)$이 $g_{1i}(\sigma_b^2)$의 불편 추정량이 아니고 $E(g_{1i}(\hat{\sigma}_b^2)) \cong g_{1i}(\sigma_b^2) - g_{3i}(\sigma_b^2)$가 성립하므로 $g_{3i}(\hat{\sigma}_b^2)$을 빼준 것이다.

Jiang과 Lahiri(1999)는 MSE(EBLUP)를 추정하는데 있어서 mse(EBLUP)의 복잡한 형식을 유도할 필요 없이 반복표본 추출법을 적용한 Jacknife방법을 이용하여 mse(EBLUP)를 계산하는 절차를 소개하였다.

Jacknife방법에서는 σ_i^2은 알고 있다고 가정하였으며 m개 소지역에서 하나의 소지역을 생략하여 추정값을 산정하여 m개의 소지역 추정값과 모든 소지역을 포함한 추정값을 이용하여 mse(EBLUP)를 계산하는데 아래와 같은 절차를 통해서 계산이 수행된다.

(절차1) 모든 표본을 이용하여 $\hat{\sigma}_b^2$과 $\hat{\beta}$을 추정하고 k번째 소지역의 자료를 제외하고 나머지 표본을 이용하여 $\hat{\sigma}_{b(k)}^2$과 $\hat{\beta}_{(k)}$를 산정한다.

(절차2) $\hat{\omega}_i = z_i^2 \hat{\sigma}_b^2 (z_i^2 \hat{\sigma}_b^2 + \sigma_i^2)^{-1}$,

$\hat{\omega}_{i(k)} = z_i^2 \hat{\sigma}_{b(k)}^2 (z_i^2 \hat{\sigma}_{b(k)}^2 + \sigma_i^2)^{-1}$,

$\hat{Y}_i^{EC} = \hat{\omega}_i \hat{Y}_i^D + (1 - \hat{\omega}_i) X_i^T \hat{\beta}$,

$\hat{Y}_{i(k)}^{EC} = \hat{\omega}_{i(k)} \hat{Y}_i^D + (1 - \hat{\omega}_{i(k)}) X_i^T \hat{\beta}_{(k)}$를 계산한다.

(절차3) $m_{1i(J)} = g_{1i}(\hat{\sigma}_b^2) - (m-1)\sum_{i=1}^{m}\{g_{1i}(\hat{\sigma}_{b(k)}^2) - g_{1i}(\hat{\sigma}_b^2)\}^2 / m$ 을 계산한다.

(절차4) $m_{2i(J)}=(m-1)\sum_{i=1}^{m}(\hat{Y}_{i(k)}^{EC}-\hat{Y}_{i}^{EC})^2/m$ 을 계산한다.

(절차5) 잭나이프 방법에 의한 EBLUP의 평균제곱오차에 대한 추정값은

$mse_{Jack}(\hat{Y}_{i}^{EC})=m_{1i(J)}+m_{2i(J)}$ 에서 산정한다.

위와 같은 절차를 통해서 계산한 $mse(EBLUP)$는 상당히 안정된 값을 갖는 특성을 갖고 있으나 계산과정이 복잡하다는 단점이 있으나 프로그램만 효율적으로 작성한다면 앞으로 많이 이용될 수 있는 방법이다.

제 7장 결 언

시군구의 실업통계 작성의 필요성은 1995년 지방자치행정제도가 시작되면서부터 예견되어 왔으나 여건 미비로 연구가 이루어지지 못하였다. 그러나 정보화 사회로 발전해 가면서 정보의 공유와 활용이 중장기 정부정책 뿐만 아니라 일반 기업의 성공여부를 결정하는 핵심요소가 되었고 따라서 정보의 인프라인 통계의 활용이 보편화 되어가면서 신속성, 정확성과 경제성을 갖춘 통계생산이 절실하게 요청되고 있다. 특히 1997년 IMF라는 환란을 거치면서 실업률이나 실업자수 등의 실업통계는 사회적인 관심사항으로 부각되었다.

국가단위 또는 시도단위의 실업통계에 대한 신속성, 정확성과 활용성을 높이는 것도 중요하지만 기초지방자치단체의 시군구 실업통계는 무엇보다도 우선 해결되어야 할 과제로 생각되었다. 조사인력과 조사비용 등의 제한으로 연구가 미진하였지만 본 연구에서 언급된 소지역 추정기법을 적용한다면 현재의 경제활동인구조사 체계에서도 일정 규모이상의 시군구에 대해서 어느 정도의 신뢰성과 정확도를 갖춘 실업통계의 생산이 가능함을 검증하였다.

소지역 추정법에서 표본설계기반의 추정기법으로서 합성 추정량과 복합 추정량을 심층적으로 연구하였으며 이들의 시군구 실업통계 작성에서 유용성을 검증하기 위해서 2001년 5월과 6월에 광주광역시와 충청북도를 대상으로 88개의 조사구를 추가로 선정하여 별도의 경제활동인구조사를 실시하였다. 이 조사된 자료에서 산정한 직접 추정량과 기존 조사구를 근거로 추계한 추정값의 상대표준오차를 비교 분석했을 때 복합 추정량에 의한 추정값이 더 바람직함을 수치적으로 검증하였다.

또한 모형기반 추정기법으로 다단수준(Multi-level)모형의 계층적 베이즈

추정량과 시계열 및 횡단면 모형의 계층적 베이즈 추정량을 소개하였으나 이들의 수치적 비교분석에서 몇몇 시군구에서 비정상적인 변동을 보여 직접 추정값의 크기의 변화에 민감한 특성을 보여주고 있어서 우리나라의 시군구 실업통계 작성에 적용하기 위해서 좀더 심층적으로 이론적 내용뿐 아니라 조사자료에 근거한 수치적 방법의 연구를 할 필요가 있다.

일반적으로 베이지안적 추정법들은 표본수가 적을 경우에는 유의한 효과를 보이지만 표본수가 일정수준 이상이 될 경우에는 효과가 크지 않은 특징이 있다.

광역시의 구단위 실업통계를 소지역 추정법인 복합 추정량을 이용한다면 조사구의 수가 구별로 10개 이상이면 추정값들이 상대표준오차를 15%이내에 들도록 할 수있고, 도의 인구 10만 이상의 시 단위에서도 시별로 조사구 수가 10개 이상이면 추정값의 상대표준오차를 20% 이내로 낮출 수 있으나 도의 10만 이하의 시와 5만 이상의 군에서는 조사구 수를 5개 이상으로 했을 때 추정값의 상대표준오차를 25% 이내로 할 수 있음을 수치적인 예를 통해서 시군구 실업통계 작성의 가능성을 입증하였다.

본 연구에서는 광주광역시와 충청북도의 2개 광역자치단체의 예를 들었으나 연구의 신뢰성을 높이기 위해서는 최소한 4-5개의 광역자치단체의 수치적인 사례연구가 필요하고, 시군구의 목표 상대허용오차를 설정하는 것도 2000년 인구주택총조사의 10%표본조사 자료를 이용하여 광범위하게 연구할 필요가 있다.

앞으로 중앙고용정보원의 구직등록 데이터베이스와 실업보험 신청자료 등의 유의성만 확인된다면 노동시장의 현황을 잘 나타내는 보조변수들을 이용하여 실업통계의 정확성과 체감성을 높일 수 있는 추정방법에 대한 연구

가 추진되어야 할 것이다.

본 연구의 기본 목적은 소지역 추정법을 이용한 시군구 실업통계 작성의 가능성을 입증함으로써 2000년 인구주택 총 조사 자료를 이용하여 2002년 경제활동인구조사를 위한 표본설계 연구의 방향을 제시하는데 있다고 할 수 있다. 따라서 차후 경제활동인구조사 표본설계에서 연구되어야 할 내용을 표본설계와 추정방법으로 나누어서 요약하였다.

표본설계에서는 표본조사구의 관리체계를 연동교체표본관리방법과 고정표본관리방법의 장단점을 비교 설명하였고 조사방법도 앞으로 조사 환경의 변화에 대비한 방안으로 방문면접조사와 전화조사(또는 인터넷조사)의 병합방법을 제시하였다.

추정방법에서는 현재의 조사구당 부여하는 가중값의 세분화된 산정 방법을 제안하였고, 해당 월의 조사된 자료만을 이용하고 있는 현재의 추정방법을 개선하여 동일한 조사구를 계속 조사한 자료들을 이용할 수 있는 시계열 모형이나 회귀모형 또는 가중평균추정법을 적용한 추정법의 연구를 언급하였으며 소지역 추정법에서 추정오차를 계산하는 방법에 대해서 앞으로 이론적 연구를 기반으로 프로그램 개발 연구까지 추진해야 함을 강조하였다.

소지역 추정법에 속한 내용은 아니지만 대규모 통계조사에서 비표본오차 관리의 한 방법으로 무응답 대체를 언급하여 앞으로 변화될 조사환경에 대비토록 하였다.

본 보고서의 연구 내용은 소지역 추정법을 이용한 시군구 실업통계 작성에 대한 기본적인 사항일 뿐이며 앞으로 일정 수준의 신뢰도를 갖춘 시군구 실업통계를 생산하기 위해서 표본설계(실업률이나 인구사회적 특성이 유사하게 모집단를 층화할 수 있는지의 여부가 소지역 추정법에서 핵심사항임),

표본조사구 관리방법(추정방법의 선정과 밀접한 관계가 있음), 보조정보(고용정보원의 데이터베이스 자료 등)이용방안과 추정기법 및 계산 프로그램들에 관한 연구가 심층적으로 추진되어야 할 것이다.

특히 소지역 추정기법을 적용하고 있는 미국이나 캐나다에서 처음에는 시범적인 적용을 통해서 문제점을 보완하고 개선 발전시켜서 전체적으로 적용하는 과정을 거쳤다.

우리나라에서도 우선 시범적으로 일부 광역시와 도에 대해서 시구와 일부 규모가 큰 군을 대상으로 소지역 추정법으로 실업통계를 작성하는 방안을 추진하여 지역통계의 발전과 정부통계의 활용성을 높이는 노력이 있어야 할 것이다.

2-3년 단시간 내에 연구결과를 평가하기보다는 심층적이고 총체적인 연구발전을 위해서 소지역 추정법에 대한 적극적이고 지속적인 지원과 협조가 필요할 것이며 앞으로 관·학간의 협동연구체제가 정착되기 위해서 폭넓은 정책적인 배려가 있어야 할 것이다.

참 고 문 헌

[1] 박홍래(2000), "개정판 통계조사론", 영지문화사

[2] 이계오외 2인(1999), "표본조사론", 한국방송대학교 출판부

[3] 통계청(1997), "가구부문 표본개편결과 보고서"

[4] 통계청(2000), "소지역 추정법 연구",통계기획국 조사관리과

[5] 통계청(2001), "캐나다 노동력조사의 방법론", 통계기획국 조사관리과

[6] 한국노동연구원(1999), "실업통계의 개선방안", 강순희, 전재식, 이계오

[8] Brisebois, F. and Mantel, H.(1996). Month-in-sample effects for the Canadian Labour Force Survey. SSC Annual Meeting, June 1996, *Proceedings of the Survey Methods Section.*

[9] Brodeur, M., Montigny, G. and Berard, H.(1995). Challenges in developing the National Longitudinal Survey of Children. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

[10] Cochran, W.G.(1977). *Sampling Techniques*, 3rd Edition, John Wiley and Sons, New York.

[11] Chen, E.J., Gambino, J., Laniel, N. and Lindeyer, J.(1994). Design and estimation issues for income in the redesign of the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

[12] Choudhry, G.H. and Rao, J.N.K.(1989). Small Area Estimation Using Models that Combine Time Series and Cross-Sectional Data. *Proceedings of the Statistics Canada Symposium on Analysis of Data in Time, October* 1989.

[13] Dufour, J., Simard, M., Allard, B. and Ray, G.(1996). Redesign of the Labour Force Survey Sample: impact on data quality. Statistics Canada, internal

document.

[14] Drew, J.D., Belanger, Y. and Foy, P.(1985). Stratification of the Canadian Labour Force Survey. *Survey Methodology*, 11, 95-110.

[15] Drew, J.D., Singh, M.P., Choudhry, G.H.(1982). Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey. Survey Methodology, 8, 17-47.

[16] Friedman, H.P. and Rubin, J.(1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62, 1159-1178.

[17] Hartley, H.O. and Rao, J.N.K.(1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.

[18] Kennedy B., Drew J.D., and Lorenz P.(1994). The Impact of Nonresponse Adjustment on Rotation Group Bias in the Canadian Labour Force Survey. Presented at the 5th International Workshop on Household Survey Nonresponse. Ottawa, Canada.

[19] Lemaitre, G.E. and Dufour J.(1987). An Integrated Method for Weighting Persons and Families. *Survey Methodology*, 13, 199-297.

[20] Lorenz, P.(1995). Labour Force Survey-Head Office Hot deck Imputation System Specifications-Version 3. Statistics Canada, internal document.

[21] Mantel, H., Laniel, N., Duval, M.C. and Marion, J.(1994). Cost modelling of alternative sample designs for rural areas in the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

[22] Mian, I.U.H. and Laniel, J.(1994). Sample allocation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

[23] Rao, J.N.K., Hartley, H.O. and Cochran, W.G.(1962). A simple procedure for unequal probability sampling without replacement. *Journal of the Royal Statistical Society, B*, 24, 482-491.

[24] Sarndal, C.E., Swensson, B. and Wretman J.(1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

[25] Sheridan, M., Drew, D. and Allard, B.(1996). Response rate and the Canadian Labour Force Survey: Luck or Good Planning? Proceedings of Statistics Canada Symposium 96 of Nonsampling Errors, 67-75.

[26] Simard, M. and Dufour, J.(1995). Impact of the introduction of Computer-Assisted Interviewing as the new Labour Force Survey data collection method. Statistics Canada, internal document.

[27] Singh, M.P., Drew, J.D., Gambino, J.G. and Mayda, F.(1990). *Methodology of the Canadian Labour Force Survey*, 1984--1990. Statistics Canada. Catalogue Number 71-526.

[28] Singh, M.P., Gambino, J. and Mantel, H.(1994). Issues and strategies for small area data (with discussion). *Survey Methodology*, 20, 3-22.

[29] Singh, A.C., Kennedy, B., Wu, S. and Brisebois, F.(1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section,* American Statistical Association.

[30] Statistics Canada(1998). Guide to the Labour Force Survey. Available on the internet at www.statcan.ca/english/concepts/labour/index.htm

[31] Sunter, D., Kinack, M., Akyeampong, E. and Charette, D.(1995). Redesigning the Canadian Labour Force Survey Questionnaire: Development and Testing. Statistics Canada internal document.

[32] Tambay, J.L. and Catlin, G.(1995). Sample Design of the National

Population Health Survey. *Health Reports*, 7, 29-38.

[33] Wolter K.M.(1985). *Introduction to Variance Estimation.* Springer-Verlag, New York.

<부록1> 광주광역시 조사구별 자료

1. 광주광역시 5월 자료

조사구	성별	경활인구	취업자	실업자	비경활인구
24001	남	8	8	0	12
24001	여	14	14	0	10
24002	남	20	18	2	8
24002	여	18	18	0	14
24003	남	20	17	3	14
24003	여	24	22	2	16
24004	남	15	15	0	7
24004	여	13	13	0	17
24005	남	20	18	2	6
24005	여	22	22	0	12
24006	남	15	13	2	4
24006	여	6	6	0	15
24007	남	13	12	1	12
24007	여	9	9	0	17
24008	남	10	10	0	11
24008	여	14	14	0	15
24009	남	12	12	0	18
24009	여	13	13	0	19
24010	남	8	7	1	7
24010	여	11	11	0	10
24011	남	16	16	0	16
24011	여	15	15	0	13
24012	남	14	14	0	7
24012	여	13	13	0	10
24013	남	17	17	0	13
24013	여	21	21	0	23
24014	남	13	11	2	11
24014	여	11	10	1	19
24015	남	15	13	2	9
24015	여	12	11	1	18
24016	남	17	17	0	15
24016	여	16	14	2	23
24017	남	16	16	0	11

24017	여	11	11	0	14
24018	남	22	21	1	11
24018	여	13	11	2	27
24019	남	15	15	0	6
24019	여	12	12	0	13
24020	남	15	15	0	11
24020	여	10	10	0	14
24021	남	18	17	1	8
24021	여	12	12	0	8
24022	남	25	25	0	11
24022	여	12	12	0	22
24023	남	20	18	2	12
24023	여	16	15	1	16
24024	남	16	16	0	7
24024	여	12	10	2	14
24025	남	23	20	3	10
24025	여	14	11	3	18
24026	남	15	14	1	3
24026	여	20	20	0	7
24027	남	17	15	2	8
24027	여	22	21	1	17
24028	남	19	19	0	8
24028	여	17	17	0	8
24029	남	15	13	2	7
24029	여	12	10	2	13
24030	남	24	21	3	7
24030	여	16	16	0	11
24031	남	14	12	2	9
24031	여	11	11	0	15
24032	남	11	11	0	9
24032	여	9	9	0	15
24033	남	15	15	0	10
24033	여	15	14	1	26
24034	남	12	12	0	15
24034	여	14	11	3	17
24035	남	17	16	1	12
24035	여	17	15	2	17
24036	남	19	18	1	4
24036	여	13	13	0	12
24037	남	17	17	0	6
24037	여	14	14	0	19
24038	남	13	13	0	10

24038	여	9	9	0	17
24039	남	16	15	1	8
24039	여	15	13	2	15
24040	남	21	20	1	10
24040	여	18	18	0	15
24041	남	10	10	0	4
24041	여	12	12	0	9
24042	남	17	17	0	10
24042	여	21	21	0	16
24043	남	21	20	1	9
24043	여	10	10	0	25
24044	남	14	14	0	7
24044	여	13	13	0	16
24045	남	11	11	0	3
24045	여	13	13	0	6
24046	남	22	20	2	10
24046	여	13	12	1	23
24047	남	10	8	2	10
24047	여	19	18	1	22
24048	남	21	21	0	6
24048	여	12	11	1	11
24049	남	17	17	0	14
24049	여	20	20	0	18
24050	남	16	15	1	10
24050	여	18	17	1	10
24051	남	14	14	0	8
24051	여	16	16	0	12
24052	남	12	12	0	7
24052	여	10	9	1	14
24053	남	13	13	0	7
24053	여	12	12	0	11
24054	남	11	10	1	7
24054	여	9	9	0	13
24055	남	21	20	1	9
24055	여	21	19	2	18
24056	남	20	19	1	6
24056	여	15	13	2	16
24057	남	22	22	0	3
24057	여	11	11	0	16
24058	남	9	8	1	11
24058	여	17	16	1	12
24059	남	8	8	0	9

24059	여	19	19	0	9
24060	남	18	18	0	3
24060	여	13	12	1	9
24061	남	20	20	0	6
24061	여	11	10	1	11
24062	남	22	22	0	5
24062	여	15	13	2	20
24063	남	23	23	0	3
24063	여	6	6	0	19
24064	남	15	14	1	11
24064	여	15	13	2	18
24065	남	19	18	1	15
24065	여	10	9	1	25
24066	남	13	13	0	14
24066	여	9	8	1	25
24067	남	17	15	2	10
24067	여	11	11	0	18
24068	남	16	15	1	5
24068	여	13	12	1	20
24069	남	8	7	1	9
24069	여	14	13	1	13
24070	남	25	24	1	8
24070	여	15	14	1	16
24071	남	21	21	0	11
24071	여	16	16	0	25
24072	남	7	7	0	9
24072	여	13	13	0	13
24073	남	21	19	2	11
24073	여	11	11	0	22
24074	남	16	15	1	4
24074	여	13	13	0	19
24075	남	23	23	0	11
24075	여	21	21	0	18
24076	남	19	19	0	6
24076	여	11	11	0	16
24077	남	21	20	1	7
24077	여	16	16	0	8
24078	남	13	10	3	2
24078	여	15	15	0	3
24079	남	22	21	1	9
24079	여	18	16	2	13
24080	남	23	21	2	0

24080	여	12	12	0	12
24081	남	27	27	0	8
24081	여	12	12	0	25
24082	남	21	21	0	2
24082	여	10	10	0	12
24083	남	17	17	0	10
24083	여	12	12	0	14
24084	남	15	15	0	4
24084	여	17	16	1	8
24086	남	20	19	1	5
24086	여	13	11	2	17
24087	남	21	21	0	4
24087	여	22	22	0	6
24088	남	22	22	0	7
24088	여	9	9	0	17
24089	남	21	20	1	2
24089	여	17	16	1	13
24090	남	24	23	1	12
24090	여	11	11	0	18
24901	남	20	20	0	14
24901	여	12	12	0	21
24902	남	18	16	2	8
24902	여	9	9	0	19
24903	남	15	14	1	9
24903	여	14	12	2	23
24904	남	10	9	1	13
24904	여	5	5	0	10
24905	남	10	9	1	11
24905	여	10	10	0	25
24906	남	13	11	2	10
24906	여	14	14	0	8
24907	남	22	21	1	5
24907	여	14	13	1	17
24908	남	17	15	2	8
24908	여	12	12	0	17
24909	남	21	18	3	11
24909	여	14	11	3	19
24910	남	23	23	0	8
24910	여	20	20	0	16
24911	남	19	19	0	9
24911	여	8	7	1	26
24912	남	19	19	0	8

24912	여	18	17	1	14
24913	남	10	9	1	14
24913	여	6	6	0	15
24914	남	13	12	1	14
24914	여	14	14	0	14

2. 광주광역시 6월 자료

조사구	성별	경활인구	취업자	실업자	비경활인구
24001	남	7	7	0	12
24001	여	14	14	0	10
24002	남	18	17	1	10
24002	여	17	17	0	14
24003	남	21	19	2	14
24003	여	26	24	2	14
24004	남	15	14	1	7
24004	여	13	13	0	18
24005	남	16	15	1	7
24005	여	21	21	0	12
24006	남	16	15	1	4
24006	여	8	8	0	13
24007	남	15	14	1	12
24007	여	11	11	0	15
24008	남	10	10	0	11
24008	여	11	11	0	17
24009	남	13	13	0	19
24009	여	13	12	1	20
24010	남	9	8	1	8
24010	여	12	12	0	10
24011	남	18	18	0	14
24011	여	15	15	0	13
24012	남	14	14	0	7
24012	여	13	13	0	9
24013	남	17	17	0	13
24013	여	21	21	0	23
24014	남	14	11	3	11
24014	여	11	10	1	17
24015	남	16	15	1	8
24015	여	13	11	2	17
24016	남	17	17	0	13
24016	여	17	16	1	22
24017	남	17	16	1	13
24017	여	10	10	0	14
24018	남	21	20	1	12
24018	여	13	11	2	25

24019	남	18	16	2	7
24019	여	16	16	0	11
24020	남	16	16	0	11
24020	여	12	11	1	13
24021	남	18	17	1	8
24021	여	12	12	0	8
24022	남	25	22	3	13
24022	여	12	12	0	22
24023	남	19	18	1	13
24023	여	15	14	1	17
24024	남	16	16	0	7
24024	여	9	8	1	16
24025	남	22	20	2	10
24025	여	14	11	3	18
24026	남	15	14	1	3
24026	여	20	20	0	7
24027	남	17	15	2	9
24027	여	24	23	1	17
24028	남	19	19	0	8
24028	여	16	16	0	8
24029	남	14	12	2	8
24029	여	10	9	1	15
24030	남	24	21	3	7
24030	여	16	16	0	11
24031	남	14	12	2	9
24031	여	10	10	0	16
24032	남	14	14	0	10
24032	여	9	9	0	16
24033	남	15	15	0	10
24033	여	15	14	1	26
24034	남	13	12	1	15
24034	여	13	11	2	19
24035	남	15	14	1	13
24035	여	15	14	1	19
24036	남	18	17	1	5
24036	여	13	13	0	12
24037	남	20	19	1	6
24037	여	17	17	0	22
24038	남	12	12	0	13
24038	여	8	8	0	18
24039	남	16	15	1	8
24039	여	16	12	4	14

24040	남	20	20	0	9
24040	여	17	17	0	17
24041	남	10	10	0	4
24041	여	12	12	0	9
24042	남	16	16	0	9
24042	여	17	17	0	16
24043	남	21	21	0	8
24043	여	9	9	0	25
24044	남	17	15	2	6
24044	여	15	13	2	17
24045	남	10	9	1	4
24045	여	13	12	1	5
24046	남	21	20	1	10
24046	여	11	11	0	25
24047	남	10	9	1	9
24047	여	21	20	1	19
24048	남	20	20	0	7
24048	여	12	11	1	11
24049	남	18	17	1	12
24049	여	20	20	0	18
24050	남	15	15	0	10
24050	여	17	16	1	9
24051	남	17	16	1	7
24051	여	17	17	0	11
24052	남	11	10	1	8
24052	여	8	7	1	16
24053	남	13	13	0	7
24053	여	12	12	0	11
24054	남	11	10	1	8
24054	여	10	9	1	13
24055	남	22	22	0	8
24055	여	22	19	3	15
24056	남	21	20	1	6
24056	여	10	10	0	21
24057	남	22	22	0	3
24057	여	12	11	1	16
24058	남	9	8	1	10
24058	여	18	17	1	10
24059	남	10	10	0	8
24059	여	18	18	0	10
24060	남	16	16	0	4
24060	여	11	10	1	10

24061	남	18	18	0	6
24061	여	11	11	0	11
24062	남	19	19	0	5
24062	여	15	15	0	21
24063	남	22	22	0	4
24063	여	9	9	0	15
24064	남	15	13	2	12
24064	여	17	16	1	18
24065	남	20	20	0	15
24065	여	11	10	1	26
24066	남	12	12	0	14
24066	여	10	9	1	25
24067	남	19	16	3	8
24067	여	13	12	1	18
24068	남	16	15	1	5
24068	여	10	10	0	23
24069	남	9	8	1	9
24069	여	14	14	0	16
24070	남	25	23	2	8
24070	여	15	14	1	16
24071	남	21	21	0	12
24071	여	17	17	0	25
24072	남	7	7	0	10
24072	여	13	13	0	13
24073	남	21	20	1	12
24073	여	14	13	1	19
24074	남	15	14	1	6
24074	여	15	14	1	16
24075	남	26	26	0	9
24075	여	20	20	0	20
24076	남	19	19	0	6
24076	여	13	13	0	13
24077	남	22	20	2	7
24077	여	15	15	0	9
24078	남	13	12	1	2
24078	여	15	15	0	3
24079	남	22	21	1	9
24079	여	19	18	1	11
24080	남	24	22	2	0
24080	여	13	12	1	13
24081	남	28	28	0	8
24081	여	13	13	0	24

24082	남	23	23	0	3
24082	여	12	12	0	13
24083	남	19	19	0	11
24083	여	14	14	0	16
24084	남	16	14	2	4
24084	여	16	16	0	7
24086	남	20	20	0	6
24086	여	14	12	2	16
24087	남	22	22	0	4
24087	여	23	23	0	6
24088	남	23	23	0	6
24088	여	9	9	0	17
24089	남	21	19	2	2
24089	여	18	18	0	12
24090	남	23	22	1	12
24090	여	14	14	0	15
24901	남	21	19	2	13
24901	여	13	13	0	20
24902	남	18	16	2	8
24902	여	8	8	0	19
24903	남	14	13	1	10
24903	여	13	13	0	25
24904	남	10	10	0	13
24904	여	5	5	0	10
24905	남	9	8	1	12
24905	여	9	8	1	26
24906	남	13	11	2	10
24906	여	13	13	0	9
24907	남	20	20	0	7
24907	여	13	13	0	18
24908	남	16	15	1	9
24908	여	14	13	1	14
24909	남	22	21	1	12
24909	여	15	12	3	19
24910	남	22	22	0	9
24910	여	17	17	0	19
24911	남	19	19	0	9
24911	여	7	6	1	25
24912	남	18	18	0	9
24912	여	16	15	1	15
24913	남	11	10	1	14
24913	여	6	6	0	16

24914	남	13	13	0	15
24914	여	14	14	0	15

<부록2> 충청북도 조사구별 자료

1. 충청북도 5월 자료

조사구	성별	경활인구	취업자	실업자	비경활인구
33001	남	13	13	0	7
33001	여	15	14	1	8
33002	남	18	16	2	16
33002	여	16	16	0	19
33003	남	10	10	0	11
33003	여	13	13	0	10
33004	남	16	15	1	13
33004	여	15	15	0	15
33005	남	22	19	3	9
33005	여	19	18	1	18
33006	남	10	9	1	10
33006	여	14	13	1	8
33007	남	14	14	0	12
33007	여	16	15	1	19
33008	남	17	15	2	4
33008	여	11	11	0	21
33009	남	18	18	0	11
33009	여	6	6	0	23
33010	남	19	18	1	3
33010	여	14	14	0	13
33011	남	21	18	3	8
33011	여	8	8	0	20
33012	남	24	24	0	5
33012	여	9	9	0	14
33013	남	18	18	0	7
33013	여	15	15	0	17
33014	남	17	17	0	6
33014	여	18	18	0	13
33015	남	6	5	1	16
33015	여	11	11	0	26
33016	남	21	20	1	13
33016	여	16	16	0	15
33017	남	15	14	1	10
33017	여	11	11	0	13

33018	남	19	17	2	9
33018	여	9	8	1	10
33019	남	15	15	0	7
33019	여	20	19	1	12
33020	남	16	15	1	9
33020	여	13	12	1	14
33021	남	23	23	0	3
33021	여	14	14	0	16
33022	남	23	23	0	3
33022	여	10	10	0	15
33023	남	22	22	0	1
33023	여	6	5	1	17
33024	남	22	22	0	6
33024	여	5	5	0	20
33027	남	13	13	0	7
33027	여	11	10	1	9
33028	남	17	16	1	5
33028	여	17	17	0	11
33029	남	18	17	1	7
33029	여	12	12	0	25
33030	남	17	17	0	7
33030	여	24	22	2	12
33031	남	20	19	1	8
33031	여	14	14	0	20
33032	남	15	15	0	9
33032	여	11	11	0	13
33033	남	17	15	2	2
33033	여	14	14	0	10
33034	남	17	17	0	4
33034	여	14	14	0	13
33035	남	12	11	1	5
33035	여	8	8	0	10
33036	남	18	18	0	4
33036	여	14	13	1	14
33037	남	16	16	0	8
33037	여	10	10	0	18
33543	남	21	20	1	4
33543	여	9	9	0	17
33545	남	31	29	2	3
33545	여	25	25	0	13
33548	남	17	17	0	12
33548	여	22	21	1	16

33552	남	19	18	1	3
33552	여	19	19	0	9
33553	남	22	22	0	2
33553	여	17	17	0	10
33557	남	20	20	0	4
33557	여	15	15	0	13
33558	남	25	23	2	12
33558	여	21	21	0	10
33723	남	21	20	1	6
33723	여	21	21	0	9
33724	남	28	28	0	9
33724	여	27	27	0	15
33725	남	25	25	0	5
33725	여	21	21	0	4
33726	남	18	18	0	2
33726	여	13	13	0	3
33738	남	23	23	0	7
33738	여	23	23	0	11
33739	남	16	15	1	7
33739	여	14	14	0	16
33740	남	18	18	0	6
33740	여	14	13	1	21
33741	남	18	18	0	7
33741	여	11	10	1	20
33742	남	16	16	0	8
33742	여	17	17	0	16
33744	남	19	19	0	0
33744	여	20	20	0	11
33746	남	22	21	1	4
33746	여	17	17	0	8
33747	남	26	26	0	8
33747	여	27	27	0	13
33749	남	15	15	0	10
33749	여	19	19	0	7
33750	남	14	14	0	5
33750	여	16	16	0	10
33751	남	25	23	2	9
33751	여	20	19	1	15
33754	남	18	18	0	4
33754	여	17	17	0	12
33755	남	11	10	1	7
33755	여	15	14	1	11

33756	남	21	21	0	5
33756	여	16	16	0	12
33759	남	15	14	1	9
33759	여	14	14	0	13
33760	남	25	25	0	4
33760	여	14	12	2	10
33761	남	11	11	0	8
33761	여	15	15	0	12
33762	남	32	32	0	6
33762	여	13	13	0	14
33763	남	30	30	0	2
33763	여	24	24	0	8
33764	남	23	23	0	4
33764	여	17	17	0	15
33901	남	12	12	0	7
33901	여	7	7	0	13
33902	남	24	24	0	5
33902	여	9	9	0	22
33903	남	27	26	1	17
33903	여	15	15	0	21
33904	남	22	22	0	9
33904	여	13	12	1	20
33905	남	17	16	1	7
33905	여	14	14	0	14
33906	남	22	21	1	5
33906	여	7	7	0	24
33907	남	23	23	0	0
33907	여	5	5	0	10
33908	남	15	15	0	12
33908	여	10	9	1	24
33909	남	14	13	1	21
33909	여	16	15	1	16
33910	남	11	11	0	8
33910	여	17	17	0	13
33911	남	21	21	0	10
33911	여	14	14	0	22
33931	남	18	18	0	11
33931	여	13	13	0	18
33932	남	20	18	2	5
33932	여	11	10	1	16
33933	남	20	17	3	5
33933	여	15	15	0	6

33934	남	21	21	0	2
33934	여	7	7	0	22
33935	남	20	18	2	10
33935	여	14	14	0	23
33936	남	21	20	1	6
33936	여	15	15	0	9
33937	남	21	19	2	5
33937	여	16	16	0	8
33938	남	18	18	0	10
33938	여	9	9	0	18
33939	남	12	12	0	12
33939	여	9	9	0	18
33940	남	20	20	0	7
33940	여	18	18	0	14
33941	남	15	14	1	6
33941	여	13	11	2	11
33942	남	26	25	1	5
33942	여	10	10	0	19
33943	남	23	23	0	7
33943	여	17	17	0	13
33944	남	22	21	1	8
33944	여	10	9	1	21
33945	남	21	19	2	7
33945	여	18	18	0	15
33946	남	13	12	1	12
33946	여	13	12	1	24
33947	남	22	20	2	7
33947	여	14	13	1	16
33948	남	22	20	2	8
33948	여	11	11	0	16
33951	남	22	22	0	7
33951	여	18	18	0	11
33952	남	16	14	2	9
33952	여	8	8	0	20
33953	남	22	22	0	0
33953	여	20	20	0	5
33954	남	14	14	0	5
33954	여	16	16	0	13
33955	남	20	19	1	7
33955	여	8	8	0	18
33956	남	16	16	0	7
33956	여	12	12	0	9

33957	남	23	21	2	9
33957	여	18	17	1	13
33958	남	25	24	1	13
33958	여	22	22	0	10
33959	남	21	20	1	7
33959	여	16	16	0	24
33960	남	19	19	0	9
33960	여	17	17	0	17
33961	남	22	21	1	2
33961	여	16	16	0	9
33962	남	25	25	0	4
33962	여	23	23	0	13
33963	남	14	14	0	5
33963	여	14	14	0	10
33964	남	25	25	0	10
33964	여	18	18	0	18
33965	남	25	25	0	6
33965	여	17	17	0	18
33966	남	19	19	0	9
33966	여	8	8	0	18
33967	남	17	17	0	1
33967	여	17	17	0	12
33968	남	15	15	0	4
33968	여	15	15	0	11
33969	남	15	15	0	1
33969	여	13	13	0	13
33970	남	17	17	0	5
33970	여	14	14	0	8
33971	남	16	16	0	6
33971	여	13	13	0	13
33972	남	22	22	0	5
33972	여	17	17	0	10
33973	남	17	16	1	6
33973	여	10	10	0	20
33974	남	19	19	0	9
33974	여	21	21	0	15
33975	남	23	23	0	6
33975	여	21	21	0	9
33976	남	14	14	0	3
33976	여	15	15	0	5
33977	남	27	27	0	5
33977	여	21	21	0	5

33978	남	20	20	0	6
33978	여	5	5	0	19
33979	남	18	18	0	8
33979	여	16	16	0	14
33980	남	26	26	0	11
33980	여	23	23	0	10
33981	남	21	20	1	5
33981	여	19	19	0	8
33982	남	11	11	0	6
33982	여	17	17	0	13
33983	남	16	15	1	7
33983	여	16	16	0	14
33984	남	20	19	1	13
33984	여	19	18	1	14
33985	남	19	19	0	1
33985	여	13	13	0	10
33986	남	20	20	0	6
33986	여	18	18	0	8
33987	남	20	20	0	4
33987	여	22	22	0	9
33988	남	19	19	0	6
33988	여	15	15	0	15
33989	남	20	20	0	6
33989	여	13	13	0	14
33990	남	19	19	0	6
33990	여	16	16	0	14
33991	남	22	22	0	2
33991	여	20	19	1	4
33992	남	17	17	0	7
33992	여	14	14	0	13
33993	남	20	20	0	4
33993	여	19	19	0	8
33994	남	20	20	0	2
33994	여	20	20	0	4
33995	남	14	14	0	8
33995	여	12	12	0	15

2. 충청북도 6월 자료

조사구	성별	경활인구	취업자	실업자	비경활인구
33001	남	12	12	0	7
33001	여	15	14	1	8
33002	남	16	15	1	18
33002	여	15	14	1	20
33003	남	12	12	0	8
33003	여	12	10	2	11
33004	남	19	17	2	14
33004	여	15	15	0	17
33005	남	24	20	4	6
33005	여	23	20	3	14
33006	남	11	10	1	9
33006	여	11	11	0	10
33007	남	14	14	0	12
33007	여	15	15	0	18
33008	남	17	15	2	4
33008	여	12	12	0	20
33009	남	18	18	0	12
33009	여	7	7	0	22
33010	남	19	18	1	3
33010	여	14	14	0	13
33011	남	22	21	1	9
33011	여	8	8	0	20
33012	남	25	25	0	4
33012	여	9	9	0	14
33013	남	19	19	0	6
33013	여	15	15	0	16
33014	남	18	18	0	6
33014	여	16	16	0	15
33015	남	9	8	1	15
33015	여	10	10	0	30
33016	남	23	22	1	11
33016	여	14	14	0	17
33017	남	16	14	2	10
33017	여	13	13	0	14
33018	남	19	17	2	9
33018	여	10	9	1	10

33019	남	17	17	0	6
33019	여	18	17	1	15
33020	남	15	15	0	9
33020	여	14	13	1	13
33021	남	21	21	0	4
33021	여	13	13	0	17
33022	남	24	24	0	3
33022	여	9	9	0	16
33023	남	22	21	1	1
33023	여	6	6	0	17
33024	남	23	23	0	5
33024	여	7	7	0	18
33027	남	14	14	0	8
33027	여	12	11	1	9
33028	남	17	16	1	5
33028	여	16	15	1	12
33029	남	17	16	1	8
33029	여	13	13	0	24
33030	남	20	19	1	5
33030	여	23	23	0	13
33031	남	21	20	1	7
33031	여	15	15	0	19
33032	남	15	15	0	11
33032	여	11	11	0	14
33033	남	18	16	2	2
33033	여	11	11	0	12
33034	남	13	13	0	5
33034	여	9	9	0	14
33035	남	12	11	1	5
33035	여	8	7	1	11
33036	남	17	17	0	4
33036	여	15	14	1	12
33037	남	17	17	0	8
33037	여	8	8	0	20
33543	남	20	20	0	5
33543	여	9	9	0	17
33545	남	31	30	1	3
33545	여	25	25	0	13
33548	남	18	18	0	13
33548	여	22	21	1	16
33552	남	19	19	0	3
33552	여	19	19	0	8

33553	남	21	21	0	3
33553	여	16	16	0	11
33557	남	19	18	1	4
33557	여	15	15	0	12
33558	남	25	24	1	12
33558	여	21	21	0	11
33723	남	21	20	1	6
33723	여	20	20	0	10
33724	남	29	29	0	9
33724	여	27	27	0	17
33725	남	26	26	0	5
33725	여	20	20	0	4
33726	남	18	18	0	3
33726	여	13	13	0	3
33738	남	22	22	0	7
33738	여	23	23	0	12
33739	남	15	14	1	9
33739	여	14	14	0	16
33740	남	17	17	0	8
33740	여	13	13	0	22
33741	남	16	16	0	7
33741	여	9	9	0	23
33742	남	16	16	0	7
33742	여	17	16	1	17
33744	남	19	19	0	1
33744	여	20	20	0	11
33746	남	21	21	0	4
33746	여	17	17	0	8
33747	남	26	26	0	8
33747	여	27	27	0	13
33749	남	16	16	0	9
33749	여	19	19	0	7
33750	남	14	14	0	5
33750	여	17	17	0	10
33751	남	25	24	1	9
33751	여	19	18	1	16
33754	남	18	18	0	3
33754	여	18	18	0	12
33755	남	11	10	1	7
33755	여	14	14	0	12
33756	남	22	22	0	5
33756	여	17	17	0	11

33759	남	15	12	3	9
33759	여	13	13	0	14
33760	남	28	27	1	4
33760	여	12	11	1	15
33761	남	12	12	0	7
33761	여	14	14	0	13
33762	남	29	29	0	5
33762	여	15	15	0	11
33763	남	30	30	0	2
33763	여	23	23	0	9
33764	남	22	22	0	4
33764	여	20	20	0	12
33901	남	12	12	0	7
33901	여	7	7	0	13
33902	남	24	24	0	5
33902	여	9	9	0	22
33903	남	29	28	1	15
33903	여	15	15	0	21
33904	남	22	22	0	9
33904	여	13	13	0	20
33905	남	17	16	1	7
33905	여	15	15	0	13
33906	남	21	21	0	6
33906	여	8	7	1	23
33907	남	22	22	0	0
33907	여	5	5	0	8
33908	남	13	12	1	12
33908	여	9	8	1	22
33909	남	14	13	1	21
33909	여	17	16	1	15
33910	남	11	11	0	7
33910	여	18	18	0	12
33911	남	22	22	0	9
33911	여	13	13	0	23
33931	남	18	18	0	11
33931	여	12	12	0	20
33932	남	21	19	2	4
33932	여	11	10	1	16
33933	남	21	17	4	3
33933	여	16	16	0	6
33934	남	21	21	0	2
33934	여	7	7	0	23

33935	남	21	20	1	10
33935	여	17	17	0	23
33936	남	20	20	0	6
33936	여	15	15	0	9
33937	남	19	19	0	7
33937	여	17	17	0	7
33938	남	16	16	0	12
33938	여	10	10	0	16
33939	남	12	12	0	12
33939	여	9	9	0	18
33940	남	20	20	0	7
33940	여	19	19	0	13
33941	남	17	16	1	6
33941	여	12	11	1	11
33942	남	27	27	0	4
33942	여	10	10	0	20
33943	남	24	24	0	6
33943	여	18	17	1	12
33944	남	22	22	0	8
33944	여	9	9	0	23
33945	남	21	19	2	7
33945	여	19	18	1	14
33946	남	13	13	0	13
33946	여	10	10	0	25
33947	남	22	20	2	7
33947	여	14	13	1	16
33948	남	22	20	2	8
33948	여	12	11	1	15
33951	남	22	22	0	7
33951	여	17	17	0	13
33952	남	17	17	0	5
33952	여	12	11	1	15
33953	남	22	22	0	0
33953	여	20	20	0	5
33954	남	14	14	0	5
33954	여	16	16	0	13
33955	남	20	18	2	7
33955	여	8	8	0	18
33956	남	16	16	0	7
33956	여	12	12	0	9
33957	남	24	21	3	8
33957	여	17	17	0	14

33958	남	29	28	1	8
33958	여	23	23	0	9
33959	남	21	20	1	7
33959	여	16	16	0	23
33960	남	20	19	1	8
33960	여	17	17	0	17
33961	남	22	21	1	2
33961	여	16	16	0	9
33962	남	25	25	0	4
33962	여	23	23	0	13
33963	남	14	14	0	5
33963	여	16	16	0	8
33964	남	26	26	0	9
33964	여	18	18	0	18
33965	남	25	25	0	6
33965	여	16	16	0	19
33966	남	18	18	0	10
33966	여	8	8	0	18
33967	남	17	17	0	1
33967	여	17	17	0	12
33968	남	15	15	0	4
33968	여	15	15	0	11
33969	남	15	15	0	1
33969	여	15	14	1	11
33970	남	20	19	1	3
33970	여	14	14	0	8
33971	남	15	15	0	6
33971	여	14	14	0	13
33972	남	22	22	0	5
33972	여	17	17	0	10
33973	남	17	16	1	5
33973	여	10	10	0	20
33974	남	19	19	0	9
33974	여	21	21	0	15
33975	남	24	23	1	6
33975	여	21	21	0	9
33976	남	14	14	0	3
33976	여	15	15	0	5
33977	남	27	27	0	5
33977	여	22	22	0	4
33978	남	19	19	0	7
33978	여	5	5	0	18

33979	남	18	18	0	9
33979	여	16	16	0	15
33980	남	26	26	0	11
33980	여	23	23	0	10
33981	남	21	20	1	6
33981	여	19	19	0	8
33982	남	11	11	0	4
33982	여	17	17	0	15
33983	남	18	17	1	7
33983	여	16	16	0	15
33984	남	22	20	2	9
33984	여	19	18	1	16
33985	남	18	18	0	1
33985	여	15	15	0	8
33986	남	21	21	0	7
33986	여	19	19	0	9
33987	남	20	20	0	4
33987	여	22	22	0	10
33988	남	20	20	0	5
33988	여	15	15	0	16
33989	남	21	21	0	5
33989	여	13	13	0	14
33990	남	21	21	0	7
33990	여	21	21	0	11
33991	남	22	22	0	2
33991	여	22	22	0	2
33992	남	18	18	0	4
33992	여	13	13	0	16
33993	남	20	20	0	5
33993	여	19	19	0	9
33994	남	20	20	0	2
33994	여	23	23	0	1
33995	남	14	14	0	8
33995	여	12	12	0	18

<부록3> 시군구 실업통계 작성 프로그램

1. 승수 계산 프로그램

(1) 계산식

sMi = sX̂i / sXi, where sX̂i

= 소지역 i의 15세이상 상주추정인구(남,여),

여기에서 sXi = 경제활동인구조사상의 15세 이상 인구

(2) 세부 알고리즘

step 1:

· 각 광역시/도의 시지역/군지역별 15세이상 상주추정인구 count

· 즉, sX̂i count

step 2:

· 경제활동인구조사 결과 파일내의 값 입력

· A = 광역시/도별 15세이상 인구수, B = 제외자수라 하면

· C = A ÷ 15세이상 전국 인구수(각 광역시/도 인구수 합) * 100

· sXi = A - (B * C)

step 3:

· 승수 최종 계산

· 승수 = step1에서 계산된 sX̂i ÷ step2에서 계산된 sXi

step 4:

· step 3에서 계산된 승수 데이터 파일에 기록

(3) 프로그램 소스 및 설명

```
/**************************
    승수 계산 프로그램
        version 1.0
```

```
**************************/
#define         MaxGroup        25
#define         MaxLine         150

#include <stdio.h>
#include <stdlib.h>

void main(void)
{
    FILE *ifp1, *ifp2, *ofp;
    float popul[2][MaxGroup], ratio[2][MaxGroup],
        counted[2][MaxGroup], total_pop[2] = {0,0}, except[2];
    char buffer[MaxLine], buf1[80], buf2[20];
    int  hang, si, cd, sex, code, i, j;

    for (j = 0; j < 2; j++)
        for (i = 0; i <= MaxGroup; i++)
            counted[j][i] = 0;
/* 각 광역시/도의 실업자를 카운트하기 위한 변수를 0으로 초기화한다 */

    // 입력 파일 열기
    if ((ifp1 = fopen("inputdat.txt", "r")) == NULL) {
        printf("Error: File not found!\n");
        exit(0);
    }
/* 데이터 파일을 열고, 만약 없으면 프로그램 종료 */

    while (fgets(buffer, MaxLine, ifp1) != NULL) {
```

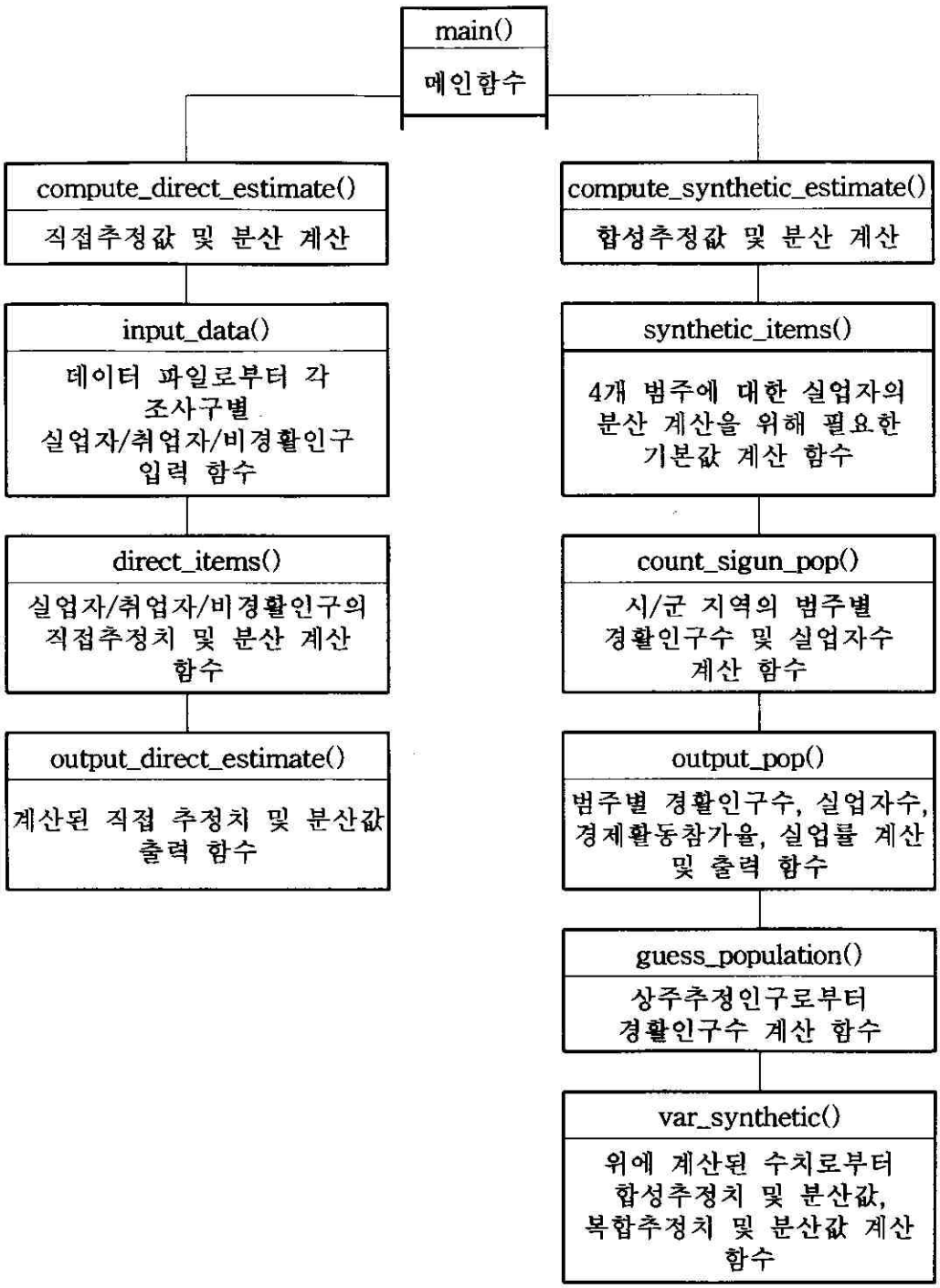
```
            sscanf(buffer, "%3d%3d%*10s%1d%*3s%1d",
                  &hang, &si, &cd, &sex);
            if (cd != 2) continue;
/* 조사 데이터에서 행정구역 코드, 조사구 번호, 성별 정보를 입력 */

            // 각 광역시/도별 코드 부여
            // 예: 31,시 -> 11, 31,군 -> 12
            if (hang == 11) code = 4;
            else if     (hang >= 21 && hang <= 26) code = hang - 16;
            else if (hang >= 31 && hang <= 39) {
                  code = (hang - 30) * 2 + 10;
                  if (si < 200 || (si >= 901 && si <= 930))
                        code--;
            }
            counted[sex-1][code-4] += 1.0;
      }
      fclose(ifp1);
/* 행정구역 코드에 대해 대응하는 첨자를 계산하고, 해당 첨자와 성별에 맞는 카
운트 변수를 1 증가 시킨다. 이를 반복함으로써 각 그룹의 실업자수를 계산할 수
있다 */

      // 추계인구 파일 열기
      if ((ifp2 = fopen("popul.txt", "r")) == NULL) {
            printf("Error: File not found!\n");
            exit(0);
      }
      fscanf(ifp2, "%f %f", except, except+1);
      for (i = 0; !feof(ifp1); i++)
```

```
            fscanf(ifp2, "%*d %f %f", &popul[0][i], &popul[1][i]);
        fclose(ifp2);
/* 추계인구 데이터 파일을 열고 각 광역시/도의 전체 추계인구수를 입력한다 */

        for (j = 0; j < 2; j++)
            for (i = 0; i <= MaxGroup; i++)
                total_pop[j] += popul[j][i];
/* 전국의 추계인구를 계산하기 위해 각 광역시/도의 추계인구수를 모두 더한다
*/

        for (j = 0; j < 2; j++)
            for (i = 0; i <= MaxGroup; i++) {
                popul[j][i] -= (popul[j][i] / total_pop[j] * 100.0) * except[j];
                if (counted[j][i] > 0)
                    ratio[j][i] = popul[j][i] / counted[j][i];
            }
/* 전국인구수에 대한 비율과 제외자수를 곱한 값을 계산하여 그 만큼 빼준 후, 이
값을 위에서 카운트한 실업자수로 나누면 해당 지역의 승수가 계산된다 */

        // 입력 파일 열고 계산된 승수 추가
        if ((ifp1 = fopen("inputdat.txt", "r")) == NULL) {
            printf("Error: File not found!\n");
            exit(0);
        }
        if ((ofp = fopen("indat.txt", "w")) == NULL) {
            printf("Error: File not found!\n");
            exit(0);
        }
```

```
    while (!feof(ifp1)) {
        fscanf(ifp1, "%74s%*7s%s", buf1, buf2);
        sscanf(buf1, "%3d%3d%*14s%1d", &hang, &si, &sex);
        if (hang == 11) code = 4;
        else if     (hang >= 21 && hang <= 26) code = hang - 16;
        else if (hang >= 31 && hang <= 39) {
            code = (hang - 30) * 2 + 10;
            if (si < 200 || (si >= 901 && si <= 930))
                code--;
        }
        fprintf(ofp, "%s%07.0f%s\n", buf1, ratio[sex-1][code-4]*1000, buf2);
    }
}
/* 계산된 승수를 데이터 파일에 첨가하여 다음 처리를 위한 준비를 하고 프로그
램을 종료한다 */
```

2. 소지역 실업자 추정 프로그램

(1) 프로그램 구성



(2) 프로그램 소스 및 설명

```
/****************************************
        경제활동인구 계산 프로그램
              version 1.5
          (소지역 추정값 및 분산 계산)
      Copyright (C) 2001 이계오/이우일
****************************************/

#define MaxLine     100
#define MaxJosagu   200

#include <stdio.h>
#include <stdlib.h>
#include <math.h>

// 전역변수 및 함수원형 선언
typedef struct count {
    int         under35;
    int         upper35;
} count;

typedef struct data {
    unsigned    josagu;
    int         sex;
    float       ratio;
    count       emp;
    count       noemp;
    count       noact;
```

} data;

/* 경제활동 인구 조사 데이터를 효율적으로 다루기 위한 자료 구조(구조체) 선언 부분으로, josagu는 조사구 코드, sex는 남/녀 구분을 위한 성별 코드(남자1, 여자 2), ratio는 해당 승수, emp는 범주별 취업자수, noemp는 범주별 실업자수, noact 는 범주별 비경활인수를 일컫는다 */

data d[2][MaxJosagu];

/* group_si_gun()는 함수는 조사구 번호순으로 나열되어 있는 데이터를 시군구 소지역으로 다시 재배열하여 출력한다. 출력된 파일을 읽어 이후에 처리되는 소지역의 실업자 추정이 이루어진다 */

// 각 시군구별 소지역으로 재배열

```
void group_si_gun(void)
{
    FILE      *ifp1, *ifp2, *ofp;

    char      src[MaxJosagu*2][MaxLine/2], buffer[MaxLine];
    unsigned  josagu, found;
    int       i = 0, index = 0, saved = -1, flag = 0;

    ifp1 = fopen("smallgrp.txt", "r");
    ifp2 = fopen("inputdat.txt", "r");
    ofp = fopen("out1.txt", "w");
```

/* *smallgrp.txt* : 각 시군구별 조사구 코드에 대한 정보 파일,
inputdat.txt : 실업자 조사 데이터 파일 */

```
    while (!feof(ifp2))
        fgets(src[index++], MaxLine-1, ifp2);
```

```
    fclose(ifp2);
/* 실업자 조사 데이터를 모두 읽어 배열 src에 저장한다 */
    while (1) {
        if (!flag) {
            i = saved + 1;
            fgets(buffer, MaxLine-1, ifp1);
            sscanf(buffer, "%*d %u", &josagu);
        }
        sscanf(src[i], "%u", &found);

        while (josagu == found && !feof(ifp1)) {
            fputs(src[i], ofp);
            fputs(src[++i], ofp);
            if (!flag) {
                saved = i;
                flag = 0;
            }
            fgets(buffer, MaxLine-1, ifp1);
            if (buffer[0] == '\n') {
                fputc('\n', ofp);
                fgets(buffer, MaxLine-1, ifp1);
            }
            sscanf(buffer, "%*d %u", &josagu);
            sscanf(src[++i], "%u", &found);
        }
        if (feof(ifp1))
            break;
```

```
        if (flag)
            i = saved + 1;
        while (josagu != found) {
            sscanf(src[i++], "%u", &found);
            i++;
        }
        flag = 1;
        i -= 2;
    }
    fclose(ifp1); fclose(ofp);
}
```

/ input_data() 함수는 광역시/도의 경제활동 조사 데이터 파일을 열고 각 지역의 승수, 범주별 취업자수, 범주별 실업자수, 범주별 비경활인구수를 읽어 온다 */*

```
// 각 시군별 실업/취업/비경 인구수를 파일로부터 읽는 함수
int input_data(FILE *ifp)
{
    char buffer[MaxLine/2];
    int  i = 0, j;

    fgets(buffer, MaxLine/2, ifp);
    while (buffer[0] != '\n' && !feof(ifp)) {
        for (j = 0; j < 2; j++) {
            sscanf(buffer, "%u %d %f %d %d %d %d %d %d",
                &d[j][i].josagu, &d[j][i].sex, &d[j][i].ratio,
                &d[j][i].emp.under35, &d[j][i].emp.upper35,
                &d[j][i].noemp.under35, &d[j][i].noemp.upper35,
```

```
                &d[j][i].noact.under35, &d[j][i].noact.upper35);
            fgets(buffer, MaxLine/2, ifp);
        }
        i++;
    }
    return i-1;
}
```

· direct_items() 함수와 output_direct_estimate() 함수 설명

step 1: 직접추정값 계산

$$\hat{Y}_{i\cdot} = \sum_s {}_s\hat{Y}_{i\cdot} \quad , \quad i=1,2,\cdots,I;\ s=1,2;\ h=1,2,\cdots,n_i$$

$$= \sum_s \sum_h {}_s\hat{Y}_{ih}$$

$$= \sum_s \sum_h {}_sM_i \ {}_sY_{ih} \ ,$$

여기에서 s= 남, 녀(1, 2),

n_i= 소지역 i의 조사구 수,

${}_sY_{ih}$ = 소지역 i의 실업자 수(남, 녀)

${}_sM_i = {}_s\hat{X}_{i\cdot} / {}_sX_{i\cdot}$; 승수(주어진 값),

${}_s\hat{X}_{i\cdot}$ = 소지역 i의 15세 이상의 상주추정인구(남, 녀),

${}_sX_{i\cdot}$ = 경제활동인구조사에서 15세 이상의 인구.

step 2: 직접추정값의 분산 계산

$$Var(\hat{Y}_{i\cdot}) = \sum_{s=1}^{2} Var({}_s\hat{Y}_{i\cdot}) + 2\,Cov({}_1\hat{Y}_{i\cdot}, {}_2\hat{Y}_{i\cdot})$$

$$= \sum_{s=1}^{2} {}_sM_i^2\, Var(\sum_{h=1}^{n_i} {}_sY_{ih}) + 2\,{}_1M_i\,{}_2M_i Cov(\sum_{h=1}^{n_i} {}_1Y_{ih}, \sum_{h=1}^{n_i} {}_2Y_{ih})$$

⇒ 분산추정공식은

$$\widehat{Var}(\hat{Y}_{i\cdot}) = \sum_{s=1}^{2} {}_sM_i^2 \left(\xi_i \sum_{h=1}^{n_i} {}_sU_{ih}^2\right) + 2\,{}_1M_i\,{}_2M_i \left(\xi_i \sum_{h=1}^{n_i} {}_1U_{ih} \cdot {}_2U_{ih}\right) \ ,$$

여기에서 $_sU_{ih} = d_sY_{ih} - {}_s\rho_i \cdot d_sX_{ih}$,

$d_sY_{ih} = {}_sY_{ih} - {}_sY_{i,h+1}$,

$d_sX_{ih} = {}_sX_{ih} - {}_sX_{i,h+1}$,

${}_s\rho_i = {}_sY_{i\cdot} / {}_sX_{i\cdot}$,

$\xi_i = [1 - n_i/(10N_i)]n_i/[2(n_i - 1)]$,

N_i =소지역 i의 모집단 조사구수.

step 3: step 1, 2에서 구한 취업자수/실업자수/비경활인구수의 직접추정치 및 분산값을 파일에 출력

```
// 취업자, 실업자, 비경제인의 추정치 및 Z, Z*Z, cov 값 계산 함수
void direct_items(FILE *ofp, int from, int to)
{
    int   i, j,
          sum_emp[2] = {0,0}, sum_noemp[2] = {0,0}, sum_noact[2] = {0,0},
          sum_emp34[2]={0,0}, sum_noemp34[2]={0,0}, sum_noact34[2]={0,0},
          emp[2][MaxJosagu], noemp[2][MaxJosagu], noact[2][MaxJosagu];
    float Zemp[2], Znoemp[2], Znoact[2], subtotal1, subtotal2, subtotal3,
          SZemp2[2] = {0,0}, SZnoemp2[2] = {0,0}, SZnoact2[2] = {0,0},
          SCovEmp = 0, SCovNoemp = 0, SCovNoact = 0;

    for (i = from; i <= to; i++) {
        for (j = 0; j < 2; j++) {
            emp[j][i] = d[j][i].emp.under35 + d[j][i].emp.upper35;
            noemp[j][i] = d[j][i].noemp.under35 + d[j][i].noemp.upper35;
            noact[j][i] = d[j][i].noact.under35 + d[j][i].noact.upper35;
            sum_emp34[j] += d[j][i].emp.under35;
            sum_noemp34[j] += d[j][i].noemp.under35;
```

```
                sum_noact34[j] += d[j][i].noact.under35;
                sum_emp[j] += emp[j][i];
                sum_noemp[j] += noemp[j][i];
                sum_noact[j] += noact[j][i];
            }
        }
/* 소지역별로 각 조사구의 취업/실업/비경활인구수와 각 수치의 합을 계산한다.
이를 이용하여 각 수치의 직접추정치 및 분산을 계산할 수 있다 */
        for (i = from+1; i <= to; i++) {
            for (j = 0; j < 2; j++) {
                subtotal1 = emp[j][i-1] + noemp[j][i-1] + noact[j][i-1];
                subtotal2 = emp[j][i] + noemp[j][i] + noact[j][i];
                subtotal1 -= subtotal2;
                subtotal3 = sum_emp[j] + sum_noemp[j] + sum_noact[j];
                Zemp[j] = emp[j][i-1] - emp[j][i] - subtotal1 *
                    sum_emp[j] / subtotal3;
                Znoemp[j] = noemp[j][i-1] - noemp[j][i] - subtotal1 *
                    sum_noemp[j] / subtotal3;
                Znoact[j] = noact[j][i-1] - noact[j][i] - subtotal1 *
                    sum_noact[j] / subtotal3;
                SZemp2[j] += Zemp[j] * Zemp[j];
                SZnoemp2[j] += Znoemp[j] * Znoemp[j];
                SZnoact2[j] += Znoact[j] * Znoact[j];
            }
            SCovEmp += Zemp[0] * Zemp[1];
            SCovNoemp += Znoemp[0] * Znoemp[1];
            SCovNoact += Znoact[0] * Znoact[1];
```

```
    }
/* 각 소지역별 취업/실업/비경활인구의 Z, Z², cov 값을 위의 공식에 의해 계산한
다 */

    for (j = 0; j < 2; j++)
    fprintf(ofp, "%5u %1d %3d %8.3f %4d %4d %4d %4d %4d %4d %4d
        %4d %4d "
        "%9.3f %9.3f %9.3f %9.3f %9.3f %9.3f\n", d[j][from].josagu,
        j+1, to-from+1, d[j][from].ratio, sum_emp[j], sum_noemp[j],
        sum_noact[j], sum_emp34[j], sum_emp[j]-sum_emp34[j],
        sum_noemp34[j], sum_noemp[j]-sum_noemp34[j], sum_noact34[j],
        sum_noact[j]-sum_noact34[j], SZemp2[j], SZnoemp2[j], SZnoact2[j],
        SCovEmp, SCovNoemp, SCovNoact);
}
/* 위에서 계산된 수치를 파일로 출력한다. 이 때 출력할 수치는 각 소지역별
조사구수, 승수, 취업자수, 실업자수, 비경활인구수, 34세이하 취업자수, 35세이상
취업자수, 34세이하 실업자수, 35세이상 실업자수, 34세이하 비경활인구수, 35세이
상 비경활인구수, Z²취업자, Z²실업자, Z²비경활인, COV취업자, COV실업자, COV비경활인
등이다 */

// 직접추정치 출력 함수
void output_direct_estimate(void)
{
    unsigned  josaguno;
    int       nofjosagu, emp[2], noemp[2], noact[2], i, j;
    float     SZemp2[2], SZnoemp2[2], SZnoact2[2], SCovEmp, SCovNoemp,
        SCovNoact,
        sum_SZemp2 = 0, sum_SZnoemp2 = 0, sum_SZnoact2 = 0,
```

```
        sum_SCovEmp = 0, sum_SCovNoemp = 0, sum_SCovNoact = 0,
            est_emp = 0, est_noemp = 0, est_noact = 0,  ratio[2], temp;
    FILE        *ifp, *ofp;

    if ((ifp = fopen("out21.txt", "r")) == NULL) {
        printf("Error: File not found!\n");
        exit(0);
    }

// 출력 파일 생성
    ofp = fopen("out3.txt", "w");
// 추정치 및 오차 계산
    while (!feof(ifp)) {
    for (i = 0; i < 2; i++) {
    fscanf(ifp, "%u %*d %d %f %d %d %d %*d %*d %*d %*d %*d %*d "
      " %f %f %f %f %f %f\n", &josaguno, &nofjosagu, &ratio[i],
      &emp[i], &noemp[i], &noact[i], &SZemp2[i], &SZnoemp2[i],
      &SZnoact2[i], &SCovEmp, &SCovNoemp, &SCovNoact);
      if (nofjosagu > 1)
      temp = ratio[i] * ratio[i] * nofjosagu / (2*(nofjosagu-1));
      else
            temp = 0;
            sum_SZemp2 += temp * SZemp2[i];
            sum_SZnoemp2 += temp * SZnoemp2[i];
            sum_SZnoact2 += temp * SZnoact2[i];
        }
        est_emp += emp[0] * ratio[0] + emp[1] * ratio[1];
        est_noemp += noemp[0] * ratio[0] + noemp[1] * ratio[1];
```

```
        est_noact += noact[0] * ratio[0] + noact[1] * ratio[1];

        if (nofjosagu > 1)
            temp = ratio[0] * ratio[1] * nofjosagu / (2*(nofjosagu-1));
        else
            temp = 0;

        sum_SCovEmp += temp * SCovEmp;
        sum_SCovNoemp += temp * SCovNoemp;
        sum_SCovNoact += temp * SCovNoact;

/* 위에 설명한 공식에 의해 실업자 추정치 및 오차를 계산하여 파일로 출력한다
*/
    if (josaguno != 33027) {
    fprintf(ofp, "%5u  %8.0f %7.0f  %8.0f %7.0f  %8.0f %7.0f\n",
        josaguno, est_emp, sqrt(sum_SZemp2 + 2 * sum_SCovEmp),
        est_noemp, sqrt(sum_SZnoemp2 + 2 * sum_SCovNoemp),
            est_noact, sqrt(sum_SZnoact2 + 2 * sum_SCovNoact));
            sum_SZemp2 = sum_SZnoemp2 = sum_SZnoact2 = 0;
            sum_SCovEmp = sum_SCovNoemp = sum_SCovNoact = 0;
            est_emp = est_noemp = est_noact = 0;
        }
    }
    fclose(ifp); fclose(ofp);
}
```

· compute direct estimate() 함수 설명

step 1: 데이터 파일(각 조사구별 경제활동 관련 설문 결과) 오픈

step 2: 시/군별로 각 범주별 취업자수, 실업자수, 비경활인구 카운트
input_data() 함수 이용

step 3: 시/군별 취업자수, 실업자수, 비경활인구의 직접추정치 계산 및 분산계산을 위한 수치값 계산
direct_items() 함수 이용

```
// 직접추정값 및 분산 계산 함수
void compute_direct_estimate(void)
{
    FILE    *ifp, *ofp;
    int     i, count;

    // 입력 파일 열기
    if ((ifp = fopen("out1.txt", "r")) == NULL) {
        printf("Error: File not found!\n");
        exit(0);
    }
    // 출력 파일 생성
    ofp = fopen("out21.txt", "w");

    while (!feof(ifp)) {
        count = input_data(ifp);
        direct_items(ofp, 0, count);
    }
    fclose(ifp); fclose(ofp);
    output_direct_estimate();
}
```

```
// 시지역/군지역의 경활인구, 실업자수를 위한 전역 변수
float siemp34[2] = {0,0}, siemp35[2] = {0,0},
        sinoemp34[2] = {0,0}, sinoemp35[2] = {0,0},
        gunemp34[2] = {0,0}, gunemp35[2] = {0,0},
        gunnoemp34[2] = {0,0}, gunnoemp35[2] = {0,0},
        siemp[2], sinoemp[2], gunemp[2], gunnoemp[2],
        sieco[2], guneco[2],
        sinoact[2] = {0,0}, gunnoact[2] = {0,0};
```

· count_sigun_pop() 함수와 output_pop() 함수 설명

step 1: 조사자료로부터 두 개의 그룹(시지역과 군지역)으로 구분하여 경제활동 인구 및 실업자수 계산

step 2: 두 그룹의 남녀 경제활동 참가율 계산

step 3: 두 그룹의 4개 범주별 실업률 계산

step 4: step 1, 2, 3에서 계산된 값을 파일에 출력

```
// 시지역 및 군지역의 그룹별 경활인구/실업자수 계산
void count_sigun_pop(int pivot, int to)
{
    int i, j;

    for (i = 0; i < pivot; i++) {
        for (j = 0; j < 2; j++) {
            siemp34[j] += d[j][i].emp.under35 * d[j][i].ratio;
            siemp35[j] += d[j][i].emp.upper35 * d[j][i].ratio;
            sinoemp34[j] += d[j][i].noemp.under35 * d[j][i].ratio;
            sinoemp35[j] += d[j][i].noemp.upper35 * d[j][i].ratio;
            sinoact[j] += (d[j][i].noact.under35+d[j][i].noact.upper35)
```

```
                    * d[j][i].ratio;
            }
        }
/* 시지역의 각 범주별 취업자/실업자/비경활인의 합계 계산 */

        for ( ; i < to; i++) {
            for (j = 0; j < 2; j++) {
                gunemp34[j] += d[j][i].emp.under35 * d[j][i].ratio;
                gunemp35[j] += d[j][i].emp.upper35 * d[j][i].ratio;
                gunnoemp34[j] += d[j][i].noemp.under35 * d[j][i].ratio;
                gunnoemp35[j] += d[j][i].noemp.upper35 * d[j][i].ratio;
                gunnoact[j] += (d[j][i].noact.under35+d[j][i].noact.upper35)
                    * d[j][i].ratio;
            }
        }
/* 군지역의 각 범주별 취업자/실업자/비경활인의 합계 계산 */

        for (j = 0; j < 2; j++) {
                siemp[j] = siemp34[j]+siemp35[j];
                sinoemp[j] = sinoemp34[j]+sinoemp35[j];
                sieco[j] = siemp[j]+sinoemp[j];
                gunemp[j] = gunemp34[j]+gunemp35[j];
                gunnoemp[j] = gunnoemp34[j]+gunnoemp35[j];
                guneco[j] = gunemp[j]+gunnoemp[j];
        }
}
/* 시지역과 군지역 각각의 취업자/실업자/비경활인수 계산
```

```
// 경활인구수, 실업자수, 경활참가율, 실업률 출력 함수
void output_pop(void)
{
    FILE *ofp;
    int         j;

    ofp = fopen("out4.txt", "w");

    // 총경활인구수|범주별 경활인구수|실업자수|경활참가율|범주별 실업률
    for (j = 0; j < 2; j++)
        fprintf(ofp, "%8.0f %8.0f %8.0f %8.0f  %7.5f  %7.5f  %7.5f\n",
            siemp[j]+sinoemp[j], siemp34[j]+sinoemp34[j],
            siemp35[j]+sinoemp35[j], sinoemp[j],
            sieco[j]/(sieco[j]+sinoact[j]),
            sinoemp34[j]/(siemp34[j]+sinoemp34[j]),
            sinoemp35[j]/(siemp35[j]+sinoemp35[j]));
/* 시지역의 총경활인구수, 범주별 경활인구수, 실업자수, 경활참가율, 범주별 실업
률 파일에 출력 */

    for (j = 0; j < 2; j++)
        fprintf(ofp, "%8.0f %8.0f %8.0f %8.0f  %7.5f  %7.5f  %7.5f\n",
            gunemp[j]+gunnoemp[j], gunemp34[j]+gunnoemp34[j],
            gunemp35[j]+gunnoemp35[j], gunnoemp[j],
            guneco[j]/(guneco[j]+gunnoact[j]),
            gunnoemp34[j]/(gunemp34[j]+gunnoemp34[j]),
            gunnoemp35[j]/(gunemp35[j]+gunnoemp35[j]));
/* 군지역의 총경활인구수, 범주별 경활인구수, 실업자수, 경활참가율, 범주별 실업
률 파일에 출력 */
```

```
    fclose(ofp);
}
```

· synthetic_items() 함수 설명

step 1: 조사구별 취업자수/실업자수/비경활인구수 데이터 파일 오픈

step 2: 시/군 두 그룹의 범주별 Z, Z^2, cov 값 계산

step 3: 계산 결과 파일에 출력

```
// 4개 범주에 대한 실업자의 Z, Z*Z, cov 값 계산 함수
void synthetic_items(void)
{
    FILE    *ifp, *ofp;
    int     i = 0, j, k, cnt = 0, mark = 0, pivot, from, to,
            countsum_emp[] = {{0,0},{0,0}}, sum_noemp[] = {{0,0},{0,0}},
            sum_noact[] = {{0,0},{0,0}};
    float   Znoemp34[2], Znoemp35[2], SZnoemp34[] = {0,0},
            SZnoemp35[] = {0,0}, SCovNoemp[] = {0,0,0,0,0,0},
            undertotal1, undertotal2, undertotal3,
            uppertotal1, uppertotal2, uppertotal3;
    char    buffer[MaxLine/2];

    // 입력 파일 열기
    if ((ifp = fopen("out1.txt", "r")) == NULL) {
        printf("Error: File not found!\n");
        exit(0);
    }
```

```
fgets(buffer, MaxLine/2, ifp);
while (!feof(ifp)) {
    for (j = 0; j < 2; j++) {
        sscanf(buffer, "%u %d %f %d %d %d %d %d %d",
            &d[j][cnt].josagu, &d[j][cnt].sex, &d[j][cnt].ratio,
            &d[j][cnt].emp.under35, &d[j][cnt].emp.upper35,
            &d[j][cnt].noemp.under35, &d[j][cnt].noemp.upper35,
            &d[j][cnt].noact.under35, &d[j][cnt].noact.upper35);
        fgets(buffer, MaxLine/2, ifp);
    }
    if (buffer[0] == '\n') {
        if (++mark == 3)
            pivot = cnt;
        fgets(buffer, MaxLine/2, ifp);
    }
    cnt++;
}
fclose(ifp);

// 출력파일 열기
ofp = fopen("out22.txt", "w");

for (k = 0; k < 2; k++) {
    from = k ? pivot + 1 : 0;
    to = k ? cnt : pivot + 1;
    for (i = from; i < to; i++) {
        for (j = 0; j < 2; j++) {
            sum_emp[j].under35 += d[j][i].emp.under35;
```

```
                sum_noemp[j].under35 += d[j][i].noemp.under35;
                sum_noact[j].under35 += d[j][i].noact.under35;
                sum_emp[j].upper35 += d[j][i].emp.upper35;
                sum_noemp[j].upper35 += d[j][i].noemp.upper35;
                sum_noact[j].upper35 += d[j][i].noact.upper35;
            }
        }

for (i = from+1; i < to; i++) {
 for (j = 0; j < 2; j++) {
      undertotal1 = d[j][i-1].noemp.under35+d[j][i-1].emp.under35
                    + d[j][i-1].noact.under35;
      undertotal2 = d[j][i].noemp.under35 + d[j][i].emp.under35
                    + d[j][i].noact.under35;
      undertotal1 -= undertotal2;
      undertotal3 = sum_emp[j].under35 + sum_noemp[j].under35
                    + sum_noact[j].under35;

      uppertotal1 = d[j][i-1].noemp.upper35 + d[j][i-1].emp.upper35
                    + d[j][i-1].noact.upper35;
      uppertotal2 = d[j][i].noemp.upper35 + d[j][i].emp.upper35
                    + d[j][i].noact.upper35;
      uppertotal1 -= uppertotal2;
      uppertotal3 = sum_emp[j].upper35 + sum_noemp[j].upper35
                    + sum_noact[j].upper35;

      Znoemp34[j] = d[j][i-1].noemp.under35-d[j][i].noemp.under35
                  - undertotal1 * sum_noemp[j].under35 / undertotal3;
```

```
        Znoemp35[j] = d[j][i-1].noemp.upper35-d[j][i].noemp.upper35
                - uppertotal1 * sum_noemp[j].upper35 / uppertotal3;

                SZnoemp34[j] += Znoemp34[j] * Znoemp34[j];
                SZnoemp35[j] += Znoemp35[j] * Znoemp35[j];
        }
        SCovNoemp[0] += Znoemp34[0] * Znoemp35[0];
        // 남35- & 남35+
        SCovNoemp[1] += Znoemp34[0] * Znoemp34[1];
         // 남35- & 여35-
        SCovNoemp[2] += Znoemp34[0] * Znoemp35[1];
         // 남35- & 여35+
        SCovNoemp[3] += Znoemp35[0] * Znoemp34[1];
         // 남35+ & 여35-
        SCovNoemp[4] += Znoemp35[0] * Znoemp35[1];
         // 남35+ & 여35+
        SCovNoemp[5] += Znoemp34[1] * Znoemp35[1];
         // 여34- & 여35+
    }
fprintf(ofp, "%d %9.3f %9.3f %9.3f %9.3f %9.3f %9.3f %9.3f %9.3f "
    "%9.3f %9.3f %9.3f %9.3f\n", to-from, d[0][to-1].ratio,
    d[1][to-1].ratio, SZnoemp34[0], SZnoemp35[0], SZnoemp34[1],
    SZnoemp35[1], SCovNoemp[0], SCovNoemp[1], SCovNoemp[2],
    SCovNoemp[3], SCovNoemp[4], SCovNoemp[5]);

    // 카운트 변수 초기화
    for (j = 0; j < 2; j++) {
        sum_emp[j].under35 = sum_emp[j].upper35 = 0;
```

```
                sum_noemp[j].under35 = sum_noemp[j].upper35 = 0;
                sum_noact[j].under35 = sum_noact[j].upper35 = 0;
                SZnoemp34[j] = SZnoemp35[j] = 0;
            }
            SCovNoemp[0] = SCovNoemp[1] = SCovNoemp[2] = 0;
            SCovNoemp[3] = SCovNoemp[4] = SCovNoemp[5] = 0;
        }
        fclose(ofp);
}
```

· guess_population() 함수 설명

step 1: 상주추정인구 데이터 입력

step 2: 각 시/군의 범주별 경제활동인구 계산

경제활동인구 = 상주추정인구 * 경제활동참가율

step 3: step2에서 계산된 값을 파일에 출력

```
// 상주추정인구로부터 경활인구 계산 함수
void guess_population(void)
{
    FILE *ifp, *ofp;
    int         count = 0;
    float pop[4], eco[4];
    char buffer[MaxLine];

    // 입력 파일(상주추정인구 데이터) 열기
    if ((ifp = fopen("guesspop.txt", "r")) == NULL) {
        printf("Error: File not found!\n");
        exit(0);
```

```
    }

    // 출력 파일 생성
    ofp = fopen("out5.txt", "w");

    fgets(buffer, MaxLine, ifp);
    do {
        //
        fscanf(ifp, "%*6s %f %f %f %f\n", pop, pop+1, pop+2, pop+3);
        if (count < 3) {
            eco[0] = pop[0] * sieco[0] / (sieco[0]+sinoact[0]);
            eco[1] = pop[1] * sieco[0] / (sieco[0]+sinoact[0]);
            eco[2] = pop[2] * sieco[1] / (sieco[1]+sinoact[1]);
            eco[3] = pop[3] * sieco[1] / (sieco[1]+sinoact[1]);
        }
        else {
            eco[0] = pop[0] * guneco[0] / (guneco[0]+gunnoact[0]);
            eco[1] = pop[1] * guneco[0] / (guneco[0]+gunnoact[0]);
            eco[2] = pop[2] * guneco[1] / (guneco[1]+gunnoact[1]);
            eco[3] = pop[3] * guneco[1] / (guneco[1]+gunnoact[1]);
        }
        count++;
        fprintf(ofp, "%8.0f %8.0f %8.0f %8.0f\n",
            eco[0], eco[1], eco[2], eco[3]);
    } while (!feof(ifp));

    fclose(ifp); fclose(ofp);
}
```

· var synthetic() 함수 설명

step 1: 합성추정값의 분산 계산

$$= \left(\frac{E1(1)}{S1}\right)^2 \left(M_1^2\zeta_1 \sum_{i=1}^{I}\sum_{h=1}^{n_{1h}} U_{1ih}^2\right)$$

$$+ \left(\frac{E1(2)}{S2}\right)^2 \left(M_2^2\zeta_2 \sum_{i=1}^{I}\sum_{h=1}^{n_{2h}} U_{2ih}^2\right)$$

$$+ \left(\frac{E1(3)}{S3}\right)^2 \left(M_3^2\zeta_3 \sum_{i=1}^{I}\sum_{h=1}^{n_{3h}} U_{3ih}^2\right)$$

$$+ \left(\frac{E1(4)}{S4}\right)^2 \left(M_4^2\zeta_4 \sum_{i=1}^{I}\sum_{h=1}^{n_{4h}} U_{4ih}^2\right)$$

$$+ 2*\Bigg[\left\{\frac{E1(1)*E1(2)}{S1*S2}\left(M_1M_2\zeta_1 \sum_{i=1}^{I}\sum_{h=1}^{n_{1h}} U_{1ih}U_{2ih}\right)\right\}$$

$$+ \left\{\frac{E1(1)*E1(3)}{S1*S3}\left(M_1M_3\zeta_1 \sum_{i=1}^{I}\sum_{h=1}^{n_{1h}} U_{1ih}U_{3ih}\right)\right\}$$

$$+ \left\{\frac{E1(1)*E1(4)}{S1*S4}\left(M_1M_4\zeta_1 \sum_{i=1}^{I}\sum_{h=1}^{n_{1h}} U_{1ih}U_{4ih}\right)\right\}$$

$$+ \left\{\frac{E1(2)*E1(3)}{S2*S3}\left(M_2M_3\zeta_2 \sum_{i=1}^{I}\sum_{h=1}^{n_{2h}} U_{2ih}U_{3ih}\right)\right\}$$

$$+ \left\{\frac{E1(2)*E1(4)}{S2*S4}\left(M_2M_4\zeta_2 \sum_{i=1}^{I}\sum_{h=1}^{n_{2h}} U_{2ih}U_{4ih}\right)\right\}$$

$$+ \left\{\frac{E1(3)*E1(4)}{S3*S4}\left(M_3M_4\zeta_3 \sum_{i=1}^{I}\sum_{h=1}^{n_{3h}} U_{3ih}U_{4ih}\right)\right\} \Bigg] ,$$

여기에서 $M_1 = M_2 \neq M_3 = M_4$: 승수,

$\zeta_j = \frac{n_j}{2(n_j-1)}$, n_j : j번째 범주의 조사구수

E1(1)...E1(4) : 범주별 경제활동인구수

S1...S4 : 시군 그룹별 경제활동인구수

step 2: 소지역(시군구)의 가중치 계산

$$\text{가중치} = \frac{\text{합성분산}}{\text{합성분산} + \text{직접추정분산}}$$

step 3: 소지역의 합성추정값 계산

합성추정값 = (범주별 경활인구수 * 범주별 실업률)의 합

step 5: 소지역의 복합추정값 계산

복합추정값 = 가중치*직접추정값 + (1-가중치)*합성추정값

step 6: 복합추정값의 표준오차 계산

표준오차 = sqrt(가중치^2*직접추정분산 + (1-가중치^2)*합성분산)

step7: CV 계산

CV = 복합추정값의 표준오차 / 복합추정값

step 8: 위에서 계산된 결과를 파일로 출력

```
// 합성추정값의 분산 계산 함수
void var_synthetic(void)
{
    FILE *ifp1, *ifp2, *ifp3, *ifp4, *ofp;
    int         josagu, count = 0, i;
    float toteco[4], eco[4], noeco[4], ratio[2], zz[4], cov[6], weight,
                synvar = 0, synest = 0, direst, dirvar, comest, error, josa;

    // 입력 파일1(시/군 범주별 경활인구 및 실업율) 열기
    if ((ifp1 = fopen("out4.txt", "r")) == NULL) {
        printf("Error: File not found!\n");
        exit(0);
    }

    // 입력 파일2(경제활동인구: E1(1)...E1(4).....E11(4)) 열기
    if ((ifp2 = fopen("out5.txt", "r")) == NULL) {
        printf("Error: File not found!\n");
        exit(0);
```

```
    }

    // 입력 파일3(Z*Z, Cov 값) 열기
    if ((ifp3 = fopen("out22.txt", "r")) == NULL) {
        printf("Error: File not found!\n");
        exit(0);
    }

    // 입력 파일4(직접추정치 및 오차) 열기
    if ((ifp4= fopen("out3.txt", "r")) == NULL) {
        printf("Error: File not found!\n");
        exit(0);
    }

    ofp = fopen("out6.txt", "w");

fscanf(ifp1, "%*f %f %f %*f %*f %f %f\n%*f %f %f %*f %*f %f %f\n",
        toteco, toteco + 1, noeco, noeco + 1, toteco + 2, toteco + 3,
        noeco + 2, noeco + 3);
fscanf(ifp3, "%d %f %f %f %f %f %f %f %f %f %f %f %f\n", &josagu,
        ratio, ratio+1, zz, zz + 1, zz + 2, zz + 3,
        cov, cov + 1, cov + 2, cov + 3, cov + 4, cov + 5);
    josa = (float)josagu / (2.0 * (josagu - 1));
    while (!feof(ifp2)) {
        fscanf(ifp2, "%f %f %f %f\n", eco, eco + 1, eco + 2, eco + 3);
        if (++count == 4) {
         fscanf(ifp1, "%*f %f %f %*f %*f %f %f\n%*f %f %f %*f %*f
                %f %f\n",
```

```
        toteco, toteco + 1, noeco, noeco + 1, toteco + 2, toteco + 3,
        noeco + 2, noeco + 3);
        fscanf(ifp3, "%d %f %f %f %f %f %f %f %f %f %f %f
              %f\n", &josagu,
              ratio, ratio+1, zz, zz + 1, zz + 2, zz + 3,
              cov, cov + 1, cov + 2, cov + 3, cov + 4, cov + 5);
        josa = (float)josagu / (2.0 * (josagu - 1));
    }
    for (i = 0; i < 4; i++) {
        synvar += (eco[i]/toteco[i]) * (eco[i]/toteco[i]) * ratio[i/2]
              * ratio[i/2] * josa * zz[i];
        synest += eco[i] * noeco[i];
    }
    synvar += 2 * ((eco[0]*eco[1]) / (toteco[0]*toteco[1]) * ratio[0] *
        ratio[0] * josa * cov[0]);
    synvar += 2 * ((eco[0]*eco[2]) / (toteco[0]*toteco[2]) * ratio[0] *
        ratio[1] * josa * cov[1]);
    synvar += 2 * ((eco[0]*eco[3]) / (toteco[0]*toteco[3]) * ratio[0] *
        ratio[1] * josa * cov[2]);
    synvar += 2 * ((eco[1]*eco[2]) / (toteco[1]*toteco[2]) * ratio[0] *
        ratio[1] * josa * cov[3]);
    synvar += 2 * ((eco[1]*eco[3]) / (toteco[1]*toteco[3]) * ratio[0] *
        ratio[1] * josa * cov[4]);
    synvar += 2 * ((eco[2]*eco[3]) / (toteco[2]*toteco[3]) * ratio[1] *
        ratio[1] * josa * cov[5]);

    fscanf(ifp4, "%*u %*f %*f %f %f %*f %*f\n", &direst, &dirvar);
dirvar *= dirvar;
```

```
        // 복합추정치 및 분산 계산 및 출력
    weight = (dirvar == 0) ? 0 : synvar / (dirvar + synvar);
    comest = weight * direst + (1 - weight) * synest;
    error = sqrt(weight*weight*dirvar + (1-weight)*(1-weight)*synvar);
    fprintf(ofp, "%10.0f %8.0f %10.0f %8.0f %10.0f %8.0f %8.3f\n",
                direst, sqrt(dirvar), synest, sqrt(synvar), comest,
                error, error / comest);
    synest = synvar = 0;
    }
    fclose(ifp1); fclose(ifp2); fclose(ifp3); fclose(ifp4);
    fclose(ofp);
}
```

· compute_synthetic_estimate() 함수 설명

step 1: 각 조사구별로 4개 범주에 대해 취업자수/실업자수/비경활인구수 입력

step 2: 시/군 그룹의 경활인구수/실업자수/경활참가율/실업률 계산 및 결과 출력
count_sigun_pop() 및 output_pop() 함수 이용

step 3: 상주추정인구 데이터 파일을 읽어 각 시군의 범주별 경제활동인구 계산
guess_population() 함수 이용

step 4: step1, 2, 3에서 계산된 값을 이용하여 실업자의 합성추정치 및 분산값 계산
synthetic_items() 함수 이용

step 5: 위에서 계산된 실업자의 직접추정치 및 분산값, 합성추정치 및 분산값을 이용하여 복합추정치 및 분산값을 계산하고 파일에 출력
var_synthetic() 함수 이용

// 합성추정값 관련 계산 함수

```
void compute_synthetic_estimate(void)
{
    FILE    *ifp, *ofp;
    int     i = 0, j;

    synthetic_items();

    if ((ifp = fopen("out21.txt", "r")) == NULL) {
        printf("Error: File not found!\n");
        exit(0);
    }

    while (!feof(ifp)) {
    for (j = 0; j < 2; j++) {
    fscanf(ifp, "%u %d %*d %f %*d %*d %*d %d %d %d %d %d %d "
                "%*f %*f %*f %*f %*f %*f\n",
                &d[j][i].josagu, &d[j][i].sex, &d[j][i].ratio,
                &d[j][i].emp.under35, &d[j][i].emp.upper35,
                &d[j][i].noemp.under35, &d[j][i].noemp.upper35,
                &d[j][i].noact.under35, &d[j][i].noact.upper35);
        }
        i++;
    }
    fclose(ifp);
    count_sigun_pop(3, i);
    output_pop();
    guess_population();
    var_synthetic();
```

```
}

// 메인 함수
void main(void)
{
        group_si_gun();
        compute_direct_estimate();
        compute_synthetic_estimate();
}
```

【소지역통계 추정법(2차 연구결과 자료)】의 내용에 관한 문의 또는 의견이 있으시면 다음 연락처를 이용하여 주시기 바랍니다.

연락처 : 「우 302-701」 대전광역시 서구 둔산동 920번지
정부대전청사 통계청 조사관리과
☎(042)481-2088~2089 Fax(042)481-2464
E-mail : kkyoung@nso.go.kr

▣ 발간에 참여한 사람들

공군사관학교 : 이 계오 교수
조사관리과장 : 김 상식
사무관 : 김 규영
담당직원 : 변 루나

인쇄 : INNO디자인 T.042-285-2601

참고사항

본 연구자료는 시·군·구 고용통계작성을 위해 소지역 통계 추정법의 학계전문가인 공군사관학교 이 계오교수에게 의뢰하여 산출된 결과물로써 연구내용은 통계청의 공식적인 입장이 아님을 밝혀둡니다.